

Fundamentals of Astrophysics

Stan Owocki

Contents

	Part I Stellar Properties	<i>page</i> 1
1	Introduction	3
	1.1 Observational vs. Physical Properties of Stars	3
	1.2 Scales and Orders of Magnitude	5
	1.3 Questions and Exercises	9
2	Inferring Astronomical Distances	10
	2.1 Angular size	10
	2.2 Trigonometric parallax	12
	2.3 Determining the Astronomical Unit (au)	15
	2.4 Solid angle	15
	2.5 Questions and Exercises	16
3	Inferring Stellar Luminosity	18
	3.1 “Standard Candle” methods for distance	18
	3.2 Intensity or Surface Brightness	19
	3.3 Apparent and absolute magnitude and the distance modulus	20
	3.4 Questions and Exercises	21
4	Inferring Surface Temperature from a Star’s Color and/or Spectrum	23
	4.1 The wave nature of light	24
	4.2 Light quanta and the Black-Body emission spectrum	24
	4.3 Inverse-temperature dependence of wavelength for peak flux	26
	4.4 Inferring stellar temperatures from photometric colors	26
	4.5 Questions and Exercises	27
5	Inferring Stellar Radius from Luminosity and Temperature	29
	5.1 Stefan-Boltzmann law for surface flux from a blackbody	29
	5.2 Questions and Exercises	30
6	Absorption Lines in Stellar Spectra	31
	6.1 Elemental composition of the Sun and stars	33

6.2	Stellar spectral type: ionization abundances as temperature diagnostic	34
6.3	Hertzsprung-Russell (H-R) diagram	35
6.4	Questions and Exercises	36
7	Surface Gravity and Escape/Orbital Speed	37
7.1	Newton's law of gravitation and stellar surface gravity	37
7.2	Surface escape speed V_{esc}	38
7.3	Speed for circular orbit	39
7.4	Virial Theorem for bound orbits	39
7.5	Questions and Exercises	40
8	Stellar Ages and Lifetimes	42
8.1	Shortness of chemical burning timescale for Sun and stars	42
8.2	Kelvin-Helmholtz timescale for gravitational contraction	42
8.3	Nuclear burning timescale	43
8.4	Age of stellar clusters from main-sequence turnoff point	44
8.5	Questions and Exercises	45
9	Inferring Stellar Space Velocities	47
9.1	Transverse speed from proper motion observations	47
9.2	Radial velocity from Doppler shift	49
9.3	Questions and Exercises	50
10	Using Binary Systems to Determine Masses and Radii	51
10.1	Visual binaries	51
10.2	Spectroscopic binaries	53
10.3	Eclipsing binaries	55
10.4	Mass-Luminosity scaling from astrometric and eclipsing binaries	56
10.5	Questions and Exercises	57
11	Inferring Stellar Rotation	59
11.1	Rotational broadening of stellar spectral lines	59
11.2	Rotational period from starspot modulation of brightness	61
11.3	Questions and Exercises	62
12	Light Intensity and Absorption	63
12.1	Intensity vs. Flux	63
12.2	Absorption mean-free-path and optical depth	65
12.3	Inter-stellar extinction and reddening	67
12.4	Questions and Exercises	68
13	Observational Methods	69
13.1	Telescopes as light buckets	69

13.2	Angular resolution	70
13.3	Space-based missions	72
13.4	Questions and Exercises	73
14	Our Sun	74
14.1	Imaging the solar disk	74
14.2	Corona and solar wind	76
14.3	Convection as a driver of solar structure and activity	78
14.4	Questions and Exercises	80
Part II	Stellar Structure & Evolution	81
15	Hydrostatic Balance between Pressure and Gravity	83
15.1	Hydrostatic equilibrium	83
15.2	Pressure scale height and thinness of surface layer	85
15.3	Hydrostatic balance in stellar interior and the virial temperature	86
15.4	Questions and Exercises	87
16	Transport of Radiation from Interior to Surface	88
16.1	Random walk of photon diffusion from stellar core to surface	88
16.2	Diffusion approximation at depth	90
16.3	Atmospheric variation of temperature with optical depth	91
16.4	Questions and Exercises	91
17	Structure of Radiative vs. Convective Stellar Envelopes	92
17.1	$L \sim M^3$ relation for hydrostatic, radiative stellar envelopes	92
17.2	Horizontal-track Kelvin-Helmholtz contraction to the main sequence	93
17.3	Convective instability and energy transport	94
17.4	Fully convective stars – the Hayashi track for proto-stellar contraction	96
18	Hydrogen Fusion and the Mass Range of Stars	98
18.1	Core temperature for H-fusion	99
18.2	Main sequence scalings for radius-mass and luminosity-temperature	100
18.3	Lower mass limit for hydrogen fusion: Brown Dwarf stars	101
18.4	Upper mass limit for stars: the Eddington Limit	102
19	Post-Main-Sequence Evolution: Low-Mass Stars	104
19.1	Core-Hydrogen burning and evolution to the Red Giant branch	105
19.2	Helium flash and core-Helium burning on the Horizontal Branch	106
19.3	Asymptotic Giant Branch to Planetary Nebula to White Dwarf	108
19.4	White Dwarf stars	108
19.5	Chandrasekhar limit for white-dwarf mass: $M < 1.4M_{\odot}$	109

20	Post-Main-Sequence Evolution: High-Mass Stars	111
	20.1 Multiple shell burning and horizontal loops in H-R diagram	111
	20.2 Core-collapse supernovae	112
	20.3 Neutron stars	114
	20.4 Black Holes	114
	20.5 Observations of stellar remnants	116
	20.6 Gravitational Waves from Merging Black Holes or Neutron Stars	118
	20.7 Questions and Exercises	121
Part III	Interstellar Medium & Formation of Stars and Planets	123
21	The Interstellar Medium	125
	21.1 Star-gas cycle	125
	21.2 Cold-Warm-Hot phases of nearly isobaric ISM	126
	21.3 Molecules and dust in cold ISM: Giant Molecular Clouds	129
	21.4 HII regions	132
	21.5 Galactic organization of ISM and star-gas interaction along spiral arms	134
22	Star Formation	136
	22.1 Jeans Criterion for gravitational contraction	136
	22.2 Cooling by molecular emission	137
	22.3 Free-fall timescale and the galactic star formation rate	138
	22.4 Fragmentation into cold cores and the Initial Mass Function (IMF)	139
	22.5 Angular momentum conservation of rotating cores and disk formation	140
	22.6 Questions and Exercises	142
23	Origin of Planetary Systems	144
	23.1 The Nebular Model	144
	23.2 Observations of Protoplanetary Disks	145
	23.3 Our Solar System	146
	23.4 The Ice Line: Gas Giants vs. Rocky Dwarfs	147
	23.5 Equilibrium Temperature	148
	23.6 Questions and Exercises	148
24	Water Planet Earth	149
	24.1 Formation of Moon by Giant Impact	149
	24.2 Water from Icy Asteroids	150
	24.3 Our Magnetic Shield	151
	24.4 Life from Oceans: Earth vs. Icy Moons	151
	24.5 Questions and Exercises	152
25	Extra-Solar Planets	153

25.1	Direct Imaging Method	153
25.2	Radial Velocity Method	154
25.3	Transit Method	155
25.4	The Exoplanet Census: 4000+ and counting	157
25.5	Search for Earth-sized Planets in the Habitable Zone	158
25.6	Questions and Exercises	159
Part IV	Our Milky Way & Other Galaxies	161
26	Our Milky Way Galaxy	163
26.1	Disk, halo, and bulge components of the Milky Way	163
26.2	Virial mass for cluster from stellar velocity dispersion inferred from Doppler shifts	166
26.3	Galactic rotation curve & dark matter	168
26.4	Super-massive black hole at the galactic center	171
27	External Galaxies	174
27.1	Cepheid variables as standard candle for distances to external galaxies	174
27.2	Galactic redshift and Hubble's law for expansion	175
27.3	Tully-Fisher Relation: $L_{gal} \propto V_{rot}^4$	177
27.4	Spiral, Elliptical, & Irregular galaxies	179
27.5	Role of Galaxy Collisions	181
28	Active Galactic Nuclei (AGNs) and Quasars	182
28.1	Basic properties and model	182
28.2	Lyman alpha clouds	183
28.3	Gravitational lensing of quasar light by foreground Galaxy Clusters	185
28.4	Gravitational redshift	187
28.5	Apparent "super-luminal" motion of quasar jets	187
29	Large Scale Structure and Eras in the Evolution of the Universe	191
29.1	Galaxy clusters & super-clusters	191
29.2	Dark matter: Hot vs. Cold, WIMPs vs. MACHOs	192
Part V	Cosmology	195
30	Newtonian Dynamical Model of Universe Expansion	197
30.1	Critical Density	197
30.2	Gravitational deceleration of increasing scale factor	198
30.2.1	Critical Universe, $\Omega_m = 1$	200
30.2.2	Closed Universe, $\Omega_m > 1$	200
30.2.3	Open Universe, $\Omega_m < 1$	201
30.3	Redshift vs. distance: Hubble law for various expansion models	201

30.4	Questions and Exercises	203
31	Accelerating Universe with a Cosmological Constant	205
31.1	White-dwarf supernova as distant standard candles	205
31.2	Cosmological Constant and Dark Energy	206
31.3	Flat Universe with Dark Energy	208
31.3.1	Exponential expansion of flat, matter-empty universe	208
31.3.2	General solutions for flat universe with dark energy	208
31.4	The “Flatness” problem	209
32	The Hot Big Bang	211
32.1	The temperature history of the universe	211
32.2	Discovery of the Cosmic Microwave Background (CMB)	212
32.3	Fluctuation Maps from COBE, WMAP, Planck	213
33	Eras in the Evolution of the Universe	216
33.1	Matter-dominated vs. Radiation-dominated eras	216
33.2	The recombination era	217
33.3	Era of nucleosynthesis	219
33.4	The particle era	220
33.5	Questions and Exercises	222
34	Cosmic inflation	223
34.1	Problems for standard Hot Big Bang model	223
34.2	The era of cosmic inflation	223
Appendix A	Atomic Energy Levels and Transitions	226
Appendix B	Equilibrium Excitation and Ionization Balance	231
Appendix C	Atomic origins of opacity	234
Appendix D	Radiative Transfer	238

Part I

Stellar Properties

1 Introduction

1.1 Observational vs. Physical Properties of Stars

What are the key physical properties we can aspire to know about a star? When we look up at the night sky, stars are just little “points of light”, but if we look carefully, we can tell that some appear brighter than others, and moreover that some have distinctly different hues or colors than others. Of course, in modern times we now know that stars are really “Suns”, with properties that are similar – within some spread – to our own Sun. They only appear much much dimmer because they are much much further away. Indeed they appear as mere “points” because they are so far away that ordinary telescopes almost never can actually resolve a distinct visible surface, the way we can resolve, even with our naked eye, that the Sun has a finite angular size.

Because we can resolve the Sun’s surface and see that it is nearly round, it is perhaps not too hard to imagine that it is a real, physical object, albeit a very special one, something we could, in principle “reach out and touch”. (Indeed a small amount of solar matter can even travel to the vicinity of the Earth through the solar wind, coronal mass ejections, and energetic particles.) As such, we can more readily imagine trying to assign values of common physical properties – e.g. distance, size, temperature, mass, age, energy emission rate, etc. – that we regularly use to characterize objects here on Earth. Of course, when we actually do so, the values we obtain dwarf anything we have direct experience with, thus stretching our imagination, and challenging the physical intuition and insights we instinctively draw upon to function in our own everyday world. But once we learn to grapple with these huge magnitudes for the Sun, we then have at our disposal that example to provide context and a relative scale to characterize other stars. And eventually as we move on to still larger scales involving stellar clusters or even whole galaxies, which might contain thousands, millions, or indeed billions of individual stars, we can try at each step to develop a relative characterization of the scales involved in these same physical quantities of size, mass, distance, etc.

So let’s consider here the properties of stars, identifying first what we can directly *observe* about a given star. Since, as we noted above, most stars are effectively a “point” source without any (easily) detectable angular extent, we might summarize what can be directly observed as three simple properties:

1. **Position on the Sky:** Once corrected for the apparent movement due to the Earth's own motion from rotation and orbiting the Sun, this can be characterized by two coordinates – analogous to latitude and longitude – on a “celestial sphere”. Before modern times, measurements of absolute position on the sky had accuracies on order an arcmin; nowadays, it is possible to get down to a few hundredths of an arcsec from ground-based telescopes, and even to about a milli-arcsec (or less in the future) from telescopes in space, where the lack of a distorting atmosphere makes images much sharper. As discussed below, the ability to measure an annual variation in the apparent position of a star due to the Earth's motion around the Sun – a phenomena known as “trigonometric parallax” – provides a key way to infer distance to at least the nearby stars.
2. **Apparent Brightness:** The ancient Greeks introduced a system by which the apparent brightness of stars is categorized in 6 bins called “magnitude”, ranging from $m = 1$ for the brightest to $m = 6$ for the dimmest visible to the naked eye. Nowadays we have instruments that can measure a star's brightness quantitatively in terms of the energy per unit area per unit time, a quantity known as the “energy flux” F , with units $\text{erg}/\text{cm}^2/\text{s}$ in CGS or W/m^2 in MKS. Because the eye is adapted to distinguish a large dynamic range of brightness, it turns out its response is *logarithmic*. And since the Greeks decided to give dimmer stars a higher magnitude, we find that magnitude scales with the log of the *inverse* flux, $m \sim \log(1/F) \sim -\log(F)$, with the $\Delta m = 5$ steps between the brightest ($m = 1$) to dimmest ($m = 6$) naked-eye star representing a *factor 100 decrease* in physical flux F . Using long exposures on large telescopes with mirrors several *meters* in diameter, we can nowadays detect individual stars with magnitudes $m > +21$, representing fluxes a million times dimmer than the limiting magnitude $m \approx +6$ visible to the naked eye.
3. **Color or “Spectrum”:** Our perception of light in three primary colors comes from the different sensitivity of receptors in our eyes to light in distinct wavelength ranges within the visible spectrum, corresponding to Red, Green, and Blue (RGB). Similarly, in astronomy, the light from a star is often passed through different sets of filters designed to transmit only light within some characteristic band of wavelengths, for example the UBV (Ultraviolet, Blue, Visible) filters that make up the so-called “Johnson photometric system”. But much more information can be gained by using a prism or (more commonly) a diffraction grating to split the light into its spectrum, defining the variation in wavelength λ of the flux, F_λ , by measuring its value within narrow wavelength bins of width $\Delta\lambda \ll \lambda$. The “spectral resolution” $\lambda/\Delta\lambda$ available depends on the instrument (spectrometer) as well as the apparent brightness of the light source, but for bright stars with modern spectrometers, the resolution can be 10,000 or more, or indeed, for the Sun, many millions. As discussed below, a key reason for seeking such high spectral resolution is to detect “spectral lines” that arise from the absorption and emission of radiation via transitions between discrete energy levels of the atoms within the star. Such spectral lines

can provide an enormous wealth of information about the composition and physical conditions in the source star.

Indeed, a key theme here is that these 3 apparently rather limited observational properties of point-stars – position, apparent brightness, and color spectrum – can, when combined with a clear understanding of some basic physical principles, allow us to infer many of the key physical properties of stars, for example:

1. **Distance**
2. **Luminosity**
3. **Temperature**
4. **Size** (i.e. Radius)
5. **Elemental Composition** (denoted as X,Y,Z for mass fraction of H, He, and of heavy “metals”)
6. **Velocity** (Both radial (toward/away) and transverse (“proper motion” across the sky))
7. **Mass** (and surface **gravity**)
8. **Age**
9. **Rotation** (Period P and/or equatorial rotation speed V_{rot})
10. **Mass loss properties** (e.g., rate \dot{M} and outflow speed V)
11. **Magnetic field**

These are ranked roughly in order of difficulty for inferring the physical property from one or more of the three types of observational data. It also roughly describes the order in which we will examine them below. In fact, except for perhaps the last two, which we will likely discuss only briefly if at all (though they happen to be two specialties of my own research), a key goal is to provide a basic understanding of the combination of physical theories, observational data, and computational methods that make it possible to infer each of the first 9 physical properties, at least for some stars.

1.2 Scales and Orders of Magnitude

Before proceeding, let us make a brief aside to discuss ways to get our heads around the enormous scales we encounter in astrophysics.

As illustrated in figure 1.1, one approach is to use a geometric progression through *powers of ten*¹, from the scale from our own bodies, which in standard metric (MKS) units is of order 1 meter (m), to the progressively larger scales in our universe.

For example, the meter itself was originally *defined* (in 1793!) as one ten millionth, or 10^{-7} , of the distance from Earth’s equator to poles; this thus means

¹ There are many online versions, including a rather dated (1977) but still informative movie titled “Powers of Ten”, which you can readily find by google; for a modern version, see <http://www.htwins.net/scale2/>.

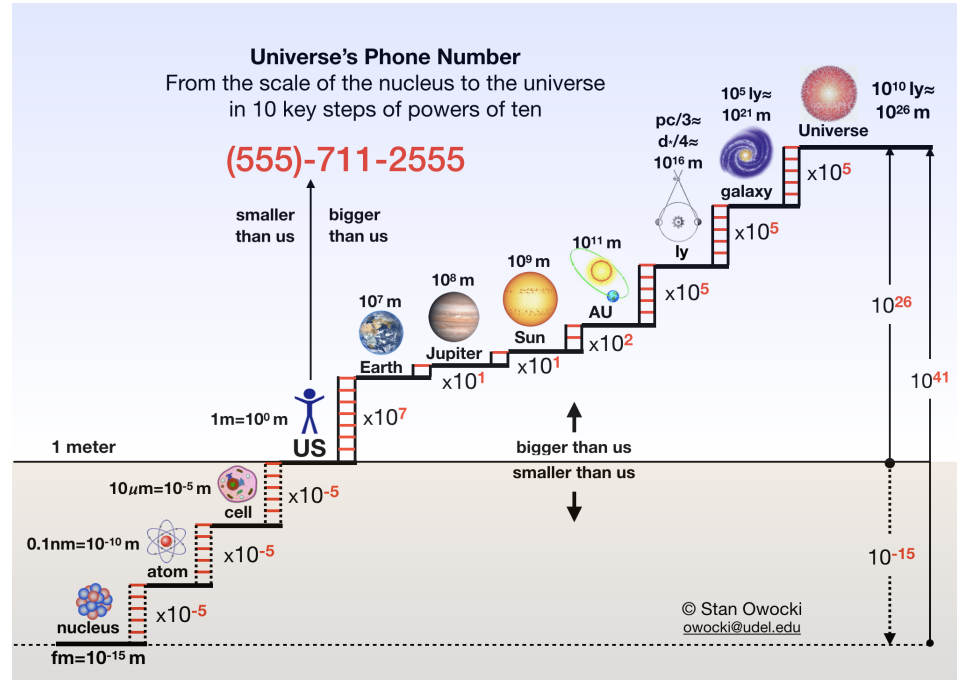


Figure 1.1 Graphic to illustrate key powers-of-ten steps between our own human scale of 1 meter, both upward to the scale of the universe (10^{26} m), and also downward to the scale of an atomic nucleus (10^{-15} m). As a mnemonic, this is cast as a 10-digit “telephone number”, with the 3-digit “area code” representing the 3 steps of 10^{-5} from us down to the nucleus, and 7-digit main-number representing 7 key steps to the scale of the universe.

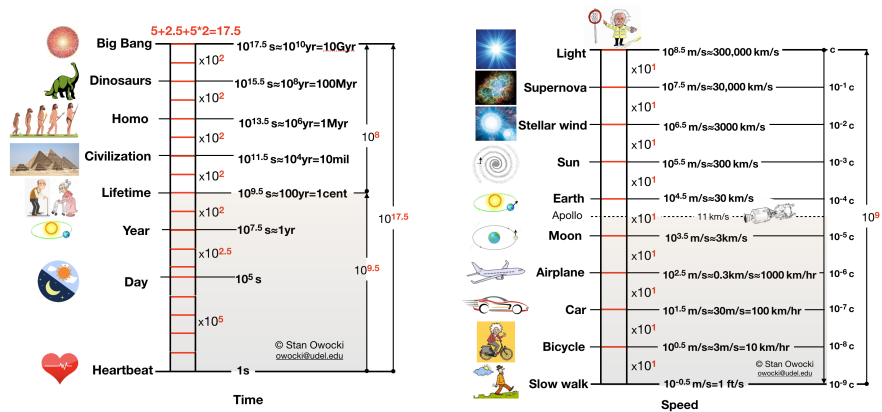


Figure 1.2 Graphics to illustrate the range of scales for time (left) and speed (right).

a total of *seven* steps in powers of ten from the scale of us to that of our Earth.

This is the largest scale for which most of us have direct experience, e.g., from overseas plane travel, or a cross country drive.

The other, rocky inner planets are somewhat smaller but same order as Earth; among the outer, gas giant planets Jupiter is the largest, about a factor ten larger than Earth, while the Sun is about another factor ten larger still, with a diameter $D_{\odot} \approx 1.4 \times 10^6$ km, about a factor hundred bigger than Earth, or of order 10^9 m.

The Earth-Sun distance, dubbed an “astronomical unit” (AU), is about about a hundred solar diameters, at 150 million km. This is of order 10^8 km = 10^{11} m, or four further powers of ten beyond the scale of our Earth, and so a total of *eleven* orders of magnitude bigger in scale than our own bodies.

An alternative way to characterize this is in terms of the time it takes light, which propagates at a speed $c = 300,000$ km/s, to reach us from the Sun; a simple calculation gives $t = d/c = 1.5e8/3e5 = 500$ s, which is about eight minutes; so we can say the Sun is 8 *light minutes* from Earth.

By contrast, it takes light from the next nearest star, Proxima Centauri, about *four years* to reach us, meaning it is at a distance of 4 *light years* (ly). A simple calculation shows that one year is $1 \text{ yr} = 365 \times 24 \times 60 \times 60 \approx 3 \times 10^7$ s; so multiplying by the speed of light $c = 3 \times 10^5$ km/s gives that $1 \text{ ly} \approx 9 \times 10^{12}$ km, or of order 10^{16} m. Thus the scale between the stars is another five order of magnitude greater than that the Earth-Sun distance, or *sixteen* orders greater than that of ourselves.

The Sun is only one of about 100 billion (10^{11}) stars in our Milky Way galaxy, a disk that is about 1000 ly thick, and about 100,000 ly across. Thus our galaxy is another five orders of magnitude bigger than the scale between individual stars, or about 10^{21} m, thus *twenty-one* orders bigger than us.

The universe itself is about 14 billion years old (14 Gyr), meaning that the most distant galaxies we can see are of order 10^{10} ly $\approx 10^{26}$ m away. We thus see that *twenty-six* powers of ten takes us from our own scale to the scale of the entire observable universe!

To recap, powers of ten steps of 7 takes us from us to the Earth; then powers of ten steps 1, 1 and 2 takes us from Earth to the size of Jupiter, Sun, and Earth-Sun distance. Then 3 successive power-ten steps of 5 take us to the distance of the nearest other star; to the size of our galaxy; and finally to the size of the universe. It can be helpful to remember this 711-2555 rule as a mnemonic – like a 7-digit telephone number – to capture the progression between key scales that characterize our place in the universe.

Indeed, we can extend this even to *small* scales, by noting that 5 powers of ten *smaller* takes us successively to the characteristic size of cell, 10^{-5} m = 10 micron; then to the size of atoms, 10^{-10} m = 0.1 nanometer; and finally to the scale of an atomic nucleus, 10 femtometer (a.k.a. “fermi”) or $1 \text{ fm} = 10^{-15}$ m.

The full sequence of steps over this span thus looks something like a 10-digit phone number with area code: 555-711-2555, representing the power of ten steps from scales of nuclei to atoms to cells to us to Earth to Jupiter to Sun to au

(distance to Sun) to light-year (\sim distance between stars) to our Galaxy to the Universe.

Finally, the enormous timescales at play in the universe can likewise be difficult to grasp.

As illustrated in the left panel of figure 1.2, humans experience time in our everyday world on the scale of a second, which is roughly the order of a single heartbeat. We live a maximum of about 100 years, or about 3 billion *seconds*. In comparison, it is estimated that the Earth is about 4.4 billion *years* old, almost as old as the Sun and the rest of the solar system. The Sun is expected to sustain its current energy output for about another 5 billion years, and so have a full lifetime of about 10 billion years. And as discussed below (see §8), the lifetimes of other stars can depend strongly on their mass; the most massive stars (about a hundred solar masses) live only about ten *million* years, while those with mass less than the Sun are expected to last for up to hundred billion years, much much longer than the current age of the universe!

The right panel of figure 1.2 gives a similar graphic for the range of speeds, from our own slow walk, through others (bicycles, cars, airplanes) we experience, then ranging to speeds of the moon, earth and Sun in their orbits, to stellar winds and supernovae, and finally ending with the maximum possible speed, the speed of light, $c = 3 \times 10^8$ m/s. The right axis relates the fraction of the light speed for each of the progression of nine powers from walking to light itself.

The remaining sections below explain how we are able to discover these fundamental properties of stars, beginning with their distance.

1.3 Questions and Exercises

Do the following computations by hand (without a calculator), to obtain results good to just one or two significant figures, but clearly showing the correct order of magnitude.

Quick Question 1:

- What speed does a person at the equator move due to Earth's rotation? Give your answer in mi/hr, km/hr, and m/s.
- What is the speed of the Earth in its orbit around the Sun? Give your answer in AU/yr, km/s, mi/hr, and in terms of the fraction of the speed of light v_{orb}/c ?
- The Sun is about 25,000 ly from the center of the Milky Way, and takes about 200 million years to complete one "Galactic year". What is the speed of Sun in its orbit around the Milky Way, in km/s. In ly/yr? In terms of the fraction of the speed of light v_{orb}/c ?

Quick Question 2: The Sun has a radius of about 700,000 km.

- How many solar radii in 1 AU? In 1 ly?
- How many Earth radii R_E in one solar radius R_\odot ?
- Solar neutrinos created in the Sun's core travel at very nearly speed of light but hardly interact with solar matter. How long does it take such core neutrinos to reach the solar surface? How long to reach us on Earth?
- What then is the solar radius in light-seconds?

Quick Question 3: The Moon is about 240,000 miles from Earth.

- What is the Earth-Moon distance in km? In light-seconds? In Earth radii R_E ? In solar radii R_\odot ?
- How many Earth-Moon distances in 1 AU?

2 Inferring Astronomical Distances

2.1 Angular size

To understand ways we might infer stellar distances, let's first consider how we intuitively estimate distance in our everyday world. Two common ways are through apparent *angular size*, and/or using our *stereoscopic vision*.

For the first, let us suppose we have some independent knowledge of the physical size of a viewed object. The apparent angular size that object subtends in our overall field of view is then used intuitively by our brains to infer the object's distance, based on our extensive experience that a greater distance makes the object subtend a smaller angle.

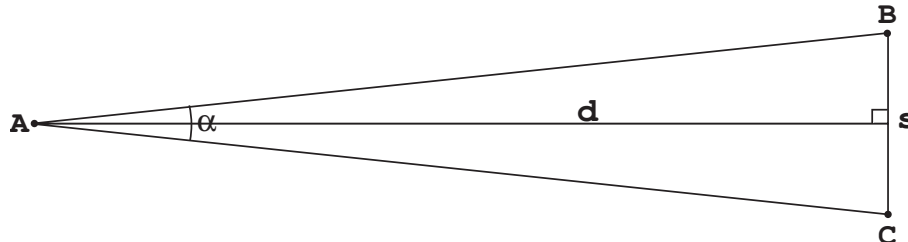


Figure 2.1 Angular size and parallax: The triangle illustrates how an object of physical size s (BC) subtends an angular size α when viewed from a point A that is at a distance d . Note that the same triangle can also illustrate the *parallax* angle α toward the point A at distance d when viewed from two points B and C separated by a length s .

As illustrated in figure 2.1, we can, with the help of some elementary geometry, formalize this intuition to write the specific formula. The triangle illustrates the angle α subtended by an object of size s from a distance d . From simple trigonometry, we find

$$\tan(\alpha/2) = \frac{s/2}{d}. \quad (2.1)$$

For distances much larger than the size, $d \gg s$, the angle is small, $\alpha \ll 1$, for which the tangent function can be approximated (e.g. by first-order Taylor expansion) to give $\tan(\alpha/2) \approx \alpha/2$, where α is measured here in radians.

(1 rad = $(180/\pi)^\circ \approx 57^\circ$). The relation between distance, size, and angle thus becomes simply

$$\boxed{\alpha \approx \frac{s}{d}}. \quad (2.2)$$

Of course, if we know the physical size and then measure the angular size, we can solve the above relation to determine the distance $d = s/\alpha$.

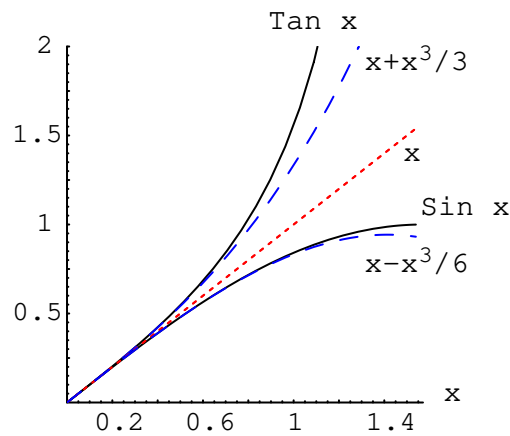


Figure 2.2 Taylor expansion of trig functions $\sin x$ and $\tan x$, about $x = 0$ to order x and order x^3 .

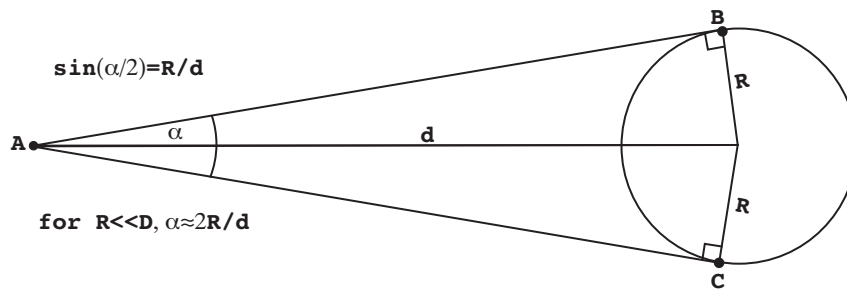


Figure 2.3 Diagram to illustrate the relation between angular size α and diameter $2R$ for a sphere at distance d .

As illustrated in figure 2.3, for a spherical object the angular size α is related to the distance d and radius R through the sine function,

$$\sin(\alpha/2) = R/d. \quad (2.3)$$

From figure 2.2 we see that, for the small angles that apply at large distances $d \gg R$, this again reduces to a simple linear form, $\alpha \approx 2R/d$, that relates size to distance.

For example, the distance from the Earth to the Sun, known as an “astronomical unit” (abbreviated “au”), is $d = 1 \text{ au} \approx 150 \times 10^6 \text{ km}$, much larger than the Sun’s physical size (i.e. diameter), which is about $s = 2R_{\odot} \approx 1.4 \times 10^6 \text{ km}$. This means that the Sun has an apparent angular diameter of

$$\alpha_{\odot} \approx \frac{2R_{\odot}}{1 \text{ au}} \approx 0.009 \text{ rad} \approx 0.5^{\circ} = 30 \text{ arcmin} = 1800 \text{ arcsec} . \quad (2.4)$$

However, as noted in §1.2 (and illustrated in figure 1.1), even the nearest stars are more than 200,000 times further away than the Sun. If we assume a similar physical radius (which actually is true for one of the components of the nearest star system, α Centauri A), then

$$\alpha_{*} = \frac{2R_{\odot}}{200,000 \text{ au}} \approx 0.009 \text{ arcsec} . \quad (2.5)$$

For ground-based telescopes, the distorting effect of the Earth’s atmosphere, known as “atmospheric seeing” (see §13.2), blurs images over an angle size of about 1 arcsec, making it very difficult to infer the actual angular size. There are some specialized techniques, e.g. “speckle interferometry”, that can just barely resolve the angular diameter of a few nearby giant stars (e.g. Betelgeuse, a.k.a. α Ori). But generally the difficulty of measuring a star’s angular size means that, even if we knew its physical size, we can not use this angular-size method to infer its distance.

2.2 Trigonometric parallax

Fortunately, there is a practical, quite direct way to infer distances to at least relatively nearby stars, namely through the method of *trigonometric parallax*.

This is physically quite analogous to the stereoscopic vision by which we use our two eyes to infer distances to objects in our everyday world. To understand this parallax effect, we can again refer to figure 2.1. If we now identify s as the *separation* between the eyes, then when we view objects at some nearby distance d , the two eyes, in order to combine the separate images as one, have to point inward an angle $\alpha = 2 \arctan(s/2d)$. Neurosensors in the eye muscles that effect this inward pointing relay this inward angle to our brain, where it is processed to provide our sense of “depth” (i.e. distance) perception.

You can easily experiment with this effect by placing your finger a few inches from your face, then blinking between your left and right eye, which thus causes the image of your finger to jump back and forth by the angle $\alpha = 2 \arctan(s/2d)$. The eye separation s is fixed, but as you move the finger closer and further away, the angle shift will become respectively larger and smaller.

Home Experiment: To illustrate this close link between parallax and angular size, try the following experiment. In front of a wall mirror, close one eye and then extend a finger from either arm to the mirror, covering the image of your closed eye. Without moving your finger, now switch the closure to the other eye. Note that the finger has

also switched to cover the other (now closed) eye, even though you didn't physically move it! Note further that this even still works as you decrease the distance from your face to the mirror. The key point here is that the “parallax” angle shift of your finger, which results from switching perspective from one eye to the other, exactly fits the apparent angular separation between your own mirror-image eyes.

Of course, for distances much more than the separation between our eyes, $d \gg s$, the angle becomes too small to perceive, and so we can only use this approach to infer distances of about, say, 10 m. But if we extend the baseline to much larger sizes s , then when coupled with accurate measures of the angle shift α , this method can be used to infer much larger distances.

For example, in the 19th century, there were efforts to use this approach to infer the distance to Mars at time when it was relatively close to Earth, namely at opposition (i.e. when Mars is on the opposite side of the Earth from the Sun). Two expeditions tried to measure the position of Mars at the same time from widely separated sites on Earth. If the distance between the sites is known, the angle difference in the measured directions to Mars, which turns out to be about an arcmin, yields a distance to Mars.

The largest separation possible from two points on the surface of the Earth is limited by the Earth's diameter. But to apply this method of trigonometric parallax to infer distances to stars, we need to use a much bigger baseline than the Earth's diameter. Fortunately though, we don't need then to go into space.

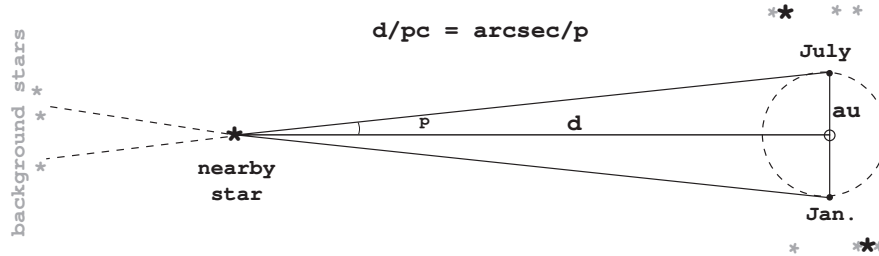


Figure 2.4 Illustration of stellar parallax, in which a relatively nearby star appears to shift against background stars by a parallax angle p as the earth moves through the 1 au radius of earth's orbit. The distance d in parsec (pc) is given by the inverse of p measured in arcsec.

As illustrated in figure 2.4, just waiting a half year from one place on the Earth allows us, as a result of the Earth's *orbit* around the Sun, to view the stars from two points separated by twice the Earth's orbital radius, i.e. 2 au. By convention, however, the associated “parallax angle” α of a star is traditionally quoted in terms of the shift from a baseline s of just *one* au. If we scale the parallax angle in units of an arcsec, the distance is

$$d = \frac{s}{\alpha} = \frac{206,265 \text{ arcsec/radian}}{\alpha/\text{radian}} \text{ au} \equiv \frac{\text{arcsec}}{\alpha} \text{ parsec}, \quad (2.6)$$

where we note that the conversion between arcsec and radian is given by $(180/\pi)$ degree/radian $\times 60$ arcmin/degree $\times 60$ arcsec/arcmin = 206,265 arcsec/radian. In the last equality, we have also introduced the distance unit *parsec* (short for “parallax second”, and often further abbreviated as “pc”), which is defined as the distance at which the parallax angle is 1 arcsec. It is thus apparent that $1 \text{ pc} = 206,265 \text{ au}$, which works out to give $1 \text{ pc} \approx 3 \times 10^{16} \text{ m}$.

The “parsec” is one of the two most common units used to characterize the huge distances we encounter in astronomy. The other is the *light-year*, which is the distance light travels in a year, at the speed of light $c = 3 \times 10^8 \text{ m/s}$. The number of seconds in a year is given by $1 \text{ yr} = 365 \times 24 \times 60 \times 60 = 3.15 \times 10^7 \text{ s}$, which, coincidentally, can be remembered as $1 \text{ yr} \approx \pi \times 10^7 \text{ s}$ (or since $\sqrt{10} \approx 3.16$, $1 \text{ yr} \approx 10^{7.5} \text{ s}$). Thus a light-year is roughly $1 \text{ ly} \approx 3\pi \times 10^{8+7} \approx 9.5 \times 10^{15} \approx 10^{16} \text{ m}$. In terms of parsecs, we can see that $1 \text{ pc} \approx 3.26 \text{ ly}$.

The parallax for even the nearest star is less than an arcsec, implying stars are all at distances more (generally *much* more) than a parsec. By repeated observation, the roughly 1 arcsec overall blurring of single stellar images by atmospheric seeing can be averaged to give a position accuracy of about $\Delta\alpha \approx 0.01 \text{ arcsec}$, implying that one can estimate distances to stars out to about $d \approx 100 \text{ pc}$. The Hipparchus satellite orbiting above Earth’s atmosphere can measure parallax angles approaching a milliarcsec ($1 \text{ mas} = 10^{-3} \text{ arcsec}$), thus potentially extending distance measurements for stars out to about a kiloparsec, $d \approx 1 \text{ kpc}$. However, parallax measurements out to such distances typically require a relatively bright source. In practice, only a fraction of all the stars (those with the highest intrinsic brightness, or “luminosity”) with distances near $d \approx 1 \text{ kpc}$ have thus far had accurate measurements of their parallax¹.

Again, from the above discussion it should be apparent that parallax is really the “flip slide” of the angular size vs. distance relation. That is, the triangle in figure 2.1 was initially used to illustrate how, from the perspective of a given point A, the angle α subtended by an object is set by the ratio of its size s to its distance d . But if we consider a simple change of observer’s perspective to the two *endpoints* (B and C) of the size segment s , then the same triangle can be used equally well to illustrate the observed parallax angle α for the point A at a distance d .

For the large ($> \text{parsec}$) distances in astronomy, it is convenient to rewrite our simple equation (2.2) to scale angular size in arcsec, with the size in au and distance in pc:

$$\boxed{\frac{\alpha}{\text{arcsec}} = \frac{s/\text{au}}{d/\text{pc}}} \quad (2.7)$$

¹ Since 2013 a follow-up satellite mission call Gaia has been in the process of measuring the absolute position and parallax to roughly one *billion* stars; see <http://sci.esa.int/gaia/>.

2.3 Determining the Astronomical Unit (au)

We thus see that determining the distance of the Earth to the Sun, i.e. measuring the physical length of an au, provides a fundamental basis for determining the distances to stars and other objects in the universe. In modern times, one way this is computed involves first measuring the distance from the Earth to the planet Venus through “radar ranging”, i.e. measuring the time Δt it takes a radar signal to bounce off Venus and return to Earth. The associated Earth-Venus distance is then given by

$$d_{EV} = \frac{c\Delta t}{2}. \quad (2.8)$$

If this distance is measured at the time when Venus has its “maximum elongation”, or maximum angular separation, from the Sun, which is found to be about 47° , then one can use simple trigonometry to derive a physical value of the au. The details are left as an exercise for the reader. (See Exercise 2-1 at the end of this section.)

2.4 Solid angle

In general objects that have a measurable angular size on the sky are extended in *two* independent directions. As the 2D generalization of an angle along just one direction, it is useful then to define for such objects a 2D *solid angle* Ω , measured now in *square radians*, but more commonly referred by the shorthand “*steradians*”.

Just as projected area A is related to the square of physical size s (or radius R), so is solid angle Ω related to the square of the *angular size* α . For an object at a distance d with projected area A , the solid angle is just

$$\Omega = \frac{A}{d^2} \approx \frac{\pi R^2}{d^2} = \pi \alpha^2, \quad (2.9)$$

where the latter equalities assume a sphere (or disk) with projected radius R and associated angular radius $\alpha = R/d$.

For more general shapes, figure 2.5 illustrates how a small solid-angle patch $\delta\Omega$ is defined in terms of ranges in the standard spherical angles representing co-latitude θ and azimuth ϕ on a sphere. An extended object would then have a solid angle given by the integral

$$\Omega = \int d\phi \sin \theta d\theta. \quad (2.10)$$

Integration over a full sphere shows that there are 4π steradians in the full sky. This represents the 2D analog to the 2π radians around the full circumference of a circle.

For our example of a circular patch of angular radius α , let us assume the

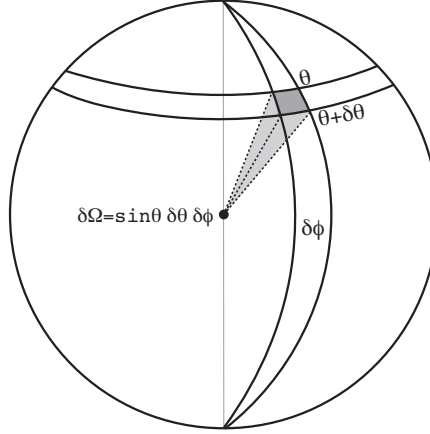


Figure 2.5 Diagram to illustrate a small patch of solid angle $\delta\Omega$ seen by an observer at the center of a sphere, with size defined by ranges in the co-latitude θ and azimuth ϕ .

object is centered around the coordinate pole – representing perhaps the image of a distant spherical object like the Sun or moon. The azimuthal symmetry means the ϕ integral evaluates to 2π , while carrying out the remaining integral over co-latitude range 0 to α then gives

$$\Omega = 2\pi [1 - \cos \alpha] . \quad (2.11)$$

In particular, applying the angular radius of the Sun $\alpha_{\odot} \approx R_{\odot}/\text{au}$ and expanding the cosine to first order (i.e., $\cos x \approx 1 - x^2/2$), we find

$$\Omega_{\odot} = 2\pi [1 - \cos(R_{\odot}/\text{au})] \approx \pi(R_{\odot}/\text{au})^2 \approx \pi\alpha_{\odot}^2 . \quad (2.12)$$

One can alternatively measure solid angle in terms of square degrees. Since there are $180/\pi \approx 57.3$ degrees in a radian, there are $(180/\pi)^2 = 57.3^2 \approx 3283$ square degrees in a steradian; the number of square degrees in the 4π steradians of the full sky is thus

$$4\pi \left(\frac{180}{\pi} \right)^2 = 41,253 \text{ deg}^2 . \quad (2.13)$$

The Sun and moon both have angular radii of about 0.25° , meaning they each have a solid angle of about $\pi(0.25)^2 = \pi/16 = 0.2 \text{ deg}^2 = 6 \times 10^{-5}$ ster, which is about 1/200,000 of the full sky².

2.5 Questions and Exercises

Quick Question 1: A helium party balloon of diameter 20 cm floats 1 meter above your head.

² If you think about it, you'll see that this helps explain why a full moon is about a million times dimmer than full sunlight! See Exercise 2-3.

- a. What is its angular diameter, in degrees and radians?
- b. What is its solid angle, in square degrees and steradians?
- c. What fraction of the full sky does it cover?
- d. At what height h would its angular diameter equal that of the Moon and Sun?

Quick Question 2:

- a. What angle α would the Earth-Sun separation subtend if viewed from a distance of $d = 1$ pc? Give your answer in both radian and arcsec.
- b. How about from a distance of $d = 1$ kpc?

Quick Question 3: Over a period of several years, two stars appear to go around each other with a fixed angular separation of 1 arcsec.

- a. What is the physical separation, in au, between the stars if they have a distance $d = 10$ pc from Earth?
- b. If they have a distance $d = 100$ pc?

Exercise 1: At the time when Venus exhibits its maximum elongation angle of about 47° from the Sun, a radar signal is found to take a round trip time $\Delta t = 667$ sec to return to Earth. Assuming both Earth and Venus have circular orbits, and using the speed of light $c = 3 \times 10^5$ km/s, compute (in km) the Earth-Sun distance, 1 AU.

Exercise 2: With a sufficiently large telescope in space, with angle error $\Delta\alpha \approx 1$ mas, for how many more stars can we expect to obtain a measured parallax than we can from ground-based surveys with $\Delta\alpha \approx 20$ mas? (Hint: What assumption do you need to make about the space density of stars in the region of the galaxy within 1 kpc from the Sun/Earth?)

Exercise 3: a. Assuming the Moon reflects a fraction a (dubbed the “albedo”) of sunlight hitting it, derive an expression for the ratio of apparent brightness (F_{moon}/F_\odot) between the full Moon and Sun, in terms of the Moon’s radius R_{moon} and its distance from earth, $d_{em} \ll \text{au}$. b. Derive the value of the albedo a for which this ratio equals the fraction of sky subtended by the Moon’s solid angle, i.e. for which $F_{\text{moon}}/F_\odot = \Omega_{\text{moon}}/4\pi$.

3 Inferring Stellar Luminosity

3.1 “Standard Candle” methods for distance

In our everyday experience, there is another way we sometimes infer distance, namely by the change in apparent brightness for objects that emit their own light, with some known power or “luminosity”. For example, a hundred watt light bulb at a distance of $d = 1$ m certainly appears a lot brighter than that same bulb at $d = 100$ m. Just as for a star, what we observe as apparent brightness is really a measure of the *flux* of light, i.e. energy per unit time *per unit area* (erg/s/cm² in CGS units, or watt/m² in MKS).

When viewing a light bulb with our eyes, it’s just the rate at which the light’s energy is captured by the area of our pupils. If we assume the light bulb’s emission is *isotropic* (i.e., the same in all directions), then as the light travels outward to a distance d , its power or luminosity is spread over a sphere of area $4\pi d^2$. This means that the light detected over a fixed detector area (like the pupil of our eye, or, for telescopes observing stars, the area of the telescope mirror) decreases in proportion to the *inverse-square* of the distance, $1/d^2$. We can thus define the apparent brightness in terms of the flux,

$$F = \frac{L}{4\pi d^2}. \quad (3.1)$$

This is a profoundly important equation in astronomy, and so you should not just memorize it, but embed it completely and deeply into your psyche.

In particular, it should become obvious that this equation can be readily used to infer the distance to an object of *known luminosity*, an approach called the *standard candle* method. (Taken from the idea that a candle, or at least a “standard” candle, has a known luminosity or intrinsic brightness.) As discussed further in sections below, there are circumstances in which we can get clues to a star’s (or other object’s) intrinsic luminosity L , for example through careful study of a star’s spectrum. If we then measure the apparent brightness (i.e. flux F), we can infer the distance through:

$$d = \sqrt{\frac{L}{4\pi F}}. \quad (3.2)$$

Indeed, when the study of a stellar spectrum is the way we infer the luminos-

ity, this method of distance determination is sometimes called “spectroscopic parallax”.

Of course, if we can independently determine the distance through the actual trigonometric parallax, then such a simple measurement of the flux can instead be used to determine the luminosity,

$$L = 4\pi d^2 F. \quad (3.3)$$

In the case of the Sun, the flux measured at Earth is referred to as the “solar constant”, with a measured mean value of about

$$F_{\odot} \approx 1.4 \frac{\text{kW}}{\text{m}^2} = 1.4 \times 10^6 \frac{\text{erg}}{\text{cm}^2 \text{s}}. \quad (3.4)$$

If we then apply the known mean distance of the Earth to the Sun, $d = 1 \text{ au}$, we obtain for the solar luminosity

$$L_{\odot} \approx 4 \times 10^{26} \text{W} = 4 \times 10^{33} \frac{\text{erg}}{\text{s}}. \quad (3.5)$$

Thus we see that the Sun emits the power of about 4×10^{24} 100-watt light bulbs! In common language this corresponds to four million billion billion, a number so huge that it loses any meaning. It illustrates again how in astronomy we have to think on a entirely different scale than we are used to in our everyday world.

But once we get used to the idea that the luminosity and other properties of the Sun are huge but still finite and measurable, we can use these as benchmarks for characterizing analogous properties of other stars and astronomical objects. In the case of stellar luminosities, for example, these typically range from about $L_{\odot}/1000$ for very cool, low-mass “dwarf” stars, to as high as $10^6 L_{\odot}$ for very hot, high-mass “supergiants”.

As discussed further below, the luminosity of a star depends directly on both its size (i.e. radius) and surface temperature. But more fundamentally these in turn are largely set by the star’s mass, age, and chemical composition.

3.2 Intensity or Surface Brightness

For any object with a resolved solid angle Ω , an important flux-related quantity is the *surface brightness* – also known as the specific intensity I ; this can be roughly (though not quite exactly; see §12.1) thought of as the *flux per solid angle*, i.e.

$$I \approx \frac{F}{\Omega} \approx \frac{L}{4\pi d^2 \pi (R/d)^2} \approx \frac{L}{4\pi^2 R^2} = \frac{F_*}{\pi}, \quad (3.6)$$

where $F_* \equiv F(R) = L/4\pi R^2$ is the *surface flux* evaluated at the stellar radius R . As illustrated in figure 3.1, the surface brightness of any resolved radiating object turns out, somewhat surprisingly, to be *independent of distance*. This is because, even though the flux declines with distance, the surface brightness ‘crowds’ this

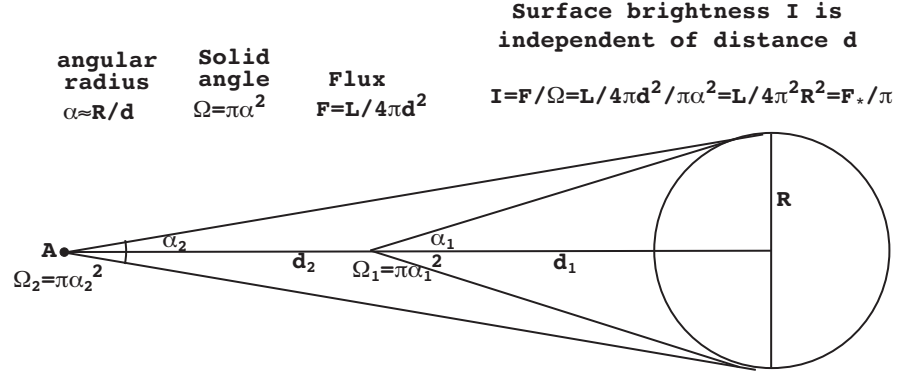


Figure 3.1 Distance independence of surface brightness of a radiating sphere, representing the flux per solid angle, $B = F/\Omega$. At greater distance d , the flux declines in proportion to $1/d^2$; but because this flux is squeezed into a smaller solid angle Ω , which also declines as $1/d^2$, the surface brightness B remains constant, independent of the distance.

flux into a proportionally smaller solid angle as the distance is increased. The ratio of flux per solid angle, or surface brightness, is thus constant.

In particular, if we ignore any absorption from earth's atmosphere, the surface brightness of the Sun that we see here on earth is actually the *same* as if we were standing on the surface of the Sun itself!

Of course, on the surface of the Sun, its radiation will fill up half the sky – i.e. 2π steradians, instead of the mere $0.2 \text{ deg}^2 = 6 \times 10^{-5}$ steradians seen from earth. The huge flux from this large, bright solid angle would cause a lot more than a mere sunburn!¹

3.3 Apparent and absolute magnitude and the distance modulus

To summarize, we have now identified 3 distinct kinds of “brightness” – absolute, apparent, and surface – associated respectively with the luminosity (energy/time), flux (energy/time/area), and specific intensity (flux emitted into a given solid angle). Before moving on to examine additional properties of stellar radiation, let us first discuss some specifics of how astronomers characterize apparent vs. absolute brightness, namely through the so-called “magnitude” system.

This system has some rather awkward conventions, developed through its long history, dating back to the ancient Greeks. As noted in §1, they ranked the apparent brightness of stars in 6 bins of magnitude, ranging from $m = 1$ for the brightest to $m = 6$ for the dimmest. Because the human eye is adapted to

¹ NASA's recently launched “Parker Solar Probe” will eventually fly within about $9R_{\odot}$ of the solar surface, or about $\sim 1/20$ au. So a key challenge has been to provide the shielding to keep the factor >400 higher solar radiation flux from frying the spacecraft's instruments.

detect a large dynamic range in brightness, it turns out that our perception of brightness depends roughly on the *logarithm* of the flux.

In our modern calibration this can be related to the Greek magnitude system by stating that a *difference of 5* in magnitude represents a *factor 100* in the relative brightness of the compared stars, with the *dimmer* star having the *larger magnitude*. This can be expressed in mathematical form as

$$m_2 - m_1 = 2.5 \log(F_1/F_2). \quad (3.7)$$

We can further extend this logarithmic magnitude system to characterize the absolute brightness, a.k.a. luminosity, of a star in terms of an *absolute* magnitude. To remove the inherent dependence on distance in the flux F , and thus in the apparent magnitude m , the absolute magnitude M is defined as the apparent magnitude that a star *would* have if it were placed at a standard distance, chosen by convention to be $d = 10$ pc. Since the flux scales with the inverse-square of distance, $F \sim 1/d^2$, the difference between apparent magnitude m and absolute magnitude M is given by

$$m - M = 5 \log(d/10 \text{ pc}), \quad (3.8)$$

which is known as the *distance modulus*.

The absolute magnitude of the Sun is $M \approx +4.8$ (though for simplicity in calculations, this is often rounded up to 5), and so the scaling for other stars can be written as

$$M = 4.8 - 2.5 \log(L/L_\odot). \quad (3.9)$$

Combining these relations, we see that the apparent magnitude of any star is given in terms of the luminosity and distance by

$$m = 4.8 - 2.5 \log(L/L_\odot) + 5 \log(d/10 \text{ pc}). \quad (3.10)$$

For bright stars, magnitudes can even become negative. For example, the (apparently) brightest star in the night sky, Sirius, has an apparent magnitude $m = -1.42$. But with a luminosity of just $L \approx 23L_\odot$, its absolute magnitude is still positive, $M = +1.40$. Its distance modulus, $m - M = -1.42 - 1.40 = -2.82$, is negative. Through eqn. (3.8), this implies that its distance, $d = 10^{1-2.82/5} = 2.7$ pc, is *less* than the standard distance of 10 pc used to define absolute magnitude and distance modulus [eqn. (3.8)].

3.4 Questions and Exercises

Quick Question 1: Recalling the relationship between an AU and a parsec from eqn. (2.6), use eqns. (3.8) and (3.9) to compute the apparent magnitude of the Sun. What then is the Sun's distance modulus?

Quick Question 2: Suppose two stars have a luminosity ratio $L_2/L_1 = 100$.

- a. At what distance ratio d_2/d_1 would the stars have the same apparent brightness, $F_2 = F_1$?
- b. For this distance ratio, what is the difference in their apparent magnitude, $m_2 - m_1$?
- c. What is the difference in their absolute magnitude, $M_2 - M_1$?
- d. What is the difference in their distance modulus?

Quick Question 3: A white-dwarf supernova with peak luminosity $L \approx 10^{10} L_\odot$ is observed to have an apparent magnitude of $m = +20$ at this peak.

- a. What is its Absolute Magnitude M ?
 - b. What is its distance d (in pc and ly). s c. How long ago did this supernova explode (in Myr)?
- (For simplicity of computation, you may take the absolute magnitude of the Sun to be $M_\odot \approx +5$.)

4 Inferring Surface Temperature from a Star's Color and/or Spectrum

Let us next consider *why* stars shine with such extreme brightness. Over the long-term (i.e., millions of years), the enormous energy emitted comes from the energy generated (by nuclear fusion) in the stellar core, as discussed further in §18 below. But the more immediate reason stars shine is more direct, namely because their surfaces are so very *hot*. The light they emit is called “thermal radiation”, and arises from the jostling of the atoms (and particularly the electrons in and around those atoms) by the violent collisions associated with the star’s high temperature¹.

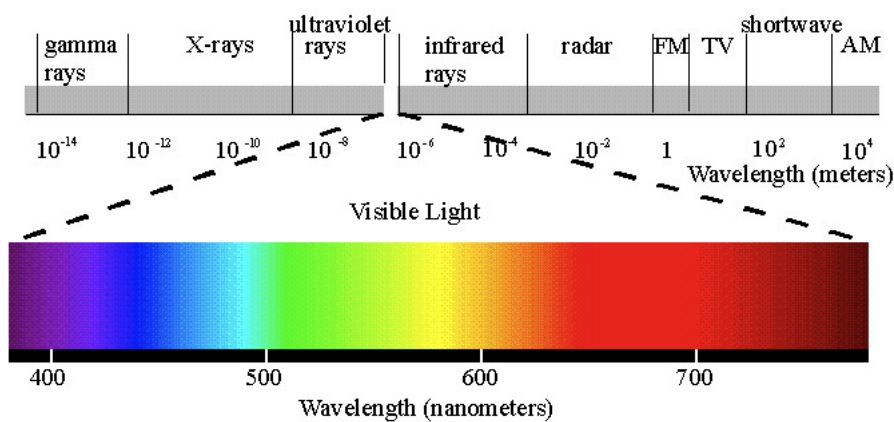


Figure 4.1 The Electromagnetic Spectrum.

¹ In astronomy, temperature is measured in a degree unit called a *Kelvin*, abbreviated *K*, and defined relative to the centigrade or “Celsius” scale *C* such that $K = C + 273$. A temperature of $T = 0\text{ K}$ is called “absolute zero”, and represents the ideal limit that all thermal motion is completely stopped. To convert from our US use of the Fahrenheit scale *F*, we first just convert to centigrade using $C = (5/9)(F - 32)$, and then add 273 to get the temperature in *K*.

4.1 The wave nature of light

To lay the groundwork for a general understanding of the key physical laws governing such thermal radiation and how it depends on temperature, we have to review what is understood about the basic nature of light, and the processes by which it is emitted and absorbed.

The 19th century physicist James Clerk Maxwell developed a set of 4 equations (Maxwell's equations) that showed how variations in Electric and Magnetic fields could lead to oscillating wave solutions, which he indeed indentified with light, or more generally *Electro-Magnetic (EM) radiation*. The wavelengths λ of these EM waves are key to their properties. As illustrated in figure 4.1, visible light corresponds to wavelengths ranging from $\lambda \approx 400$ nm (violet) to $\lambda \approx 750$ nm (red), but the full spectrum extends much further, including Ultra-Violet (UV), X-rays, and gamma rays at shorter wavelengths, and InfraRed (IR), microwaves, and radio waves at longer wavelengths. White light is made up of a broad mix of visible light ranging from Red through Green to Blue (RGB).

In a vacuum, all these EM waves travel at the *same speed*, namely the speed of light, customarily denoted as c , with a value $c \approx 3 \times 10^5$ km/s $= 3 \times 10^8$ m/s $= 3 \times 10^{10}$ cm/s. The wave *period* is the time it takes for a complete wavelength to pass a fixed point at this speed, and so is given by $P = \lambda/c$. We can thus see that the sequence of wave crests passes by at a *frequency* of once per period, $\nu = 1/P$, implying a simple relationship between light's wavelength λ , frequency ν , and speed c ,

$$\boxed{\frac{\lambda}{P} = \lambda\nu = c} . \quad (4.1)$$

4.2 Light quanta and the Black-Body emission spectrum

The wave nature of light has been confirmed by a wide range experiments. However, at the beginning of the 20th century, work by Einstein, Planck, and others led to the realization that light waves are also *quantized* into discrete wave “bundles” called *photons*. Each photon carries a discrete, indivisible “quantum” of energy that depends on the wave frequency as

$$\boxed{E = h\nu} , \quad (4.2)$$

where h is *Planck's constant*, with value $h \approx 6.6 \times 10^{-27}$ erg s $= 6.6 \times 10^{-34}$ Joule s.

This quantization of light (and indeed of all energy) has profound and wide-ranging consequences, most notably in the current context for how thermally emitted radiation is distributed in wavelength or frequency. This is known as the “Spectral Energy Distribution” (SED). For a so-called *Black Body* – meaning

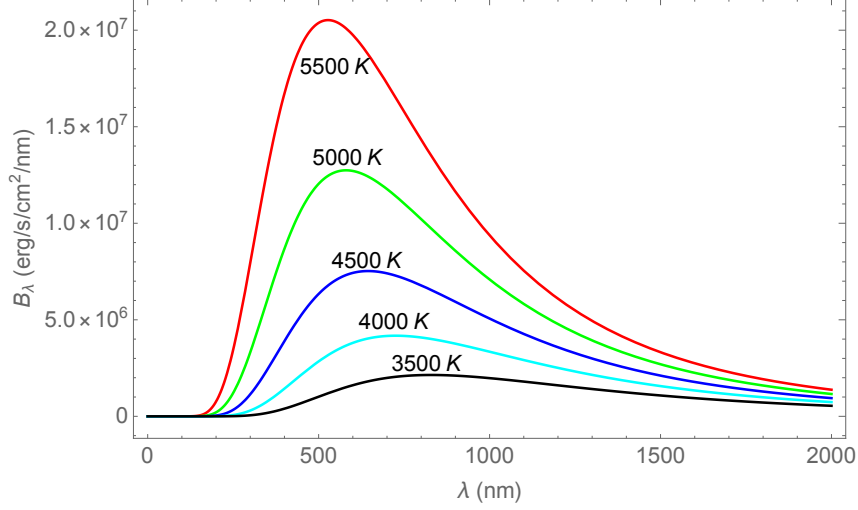


Figure 4.2 The Planck Black-Body Spectral Energy Distribution (SED) vs. wavelength λ , plotted for various temperatures T .

idealized material that is readily able to absorb and emit radiation of all wavelengths –, Planck showed that as thermal motions of the material approach a *Thermodynamic Equilibrium* (TE) in the exchange of energy between radiation and matter, the SED can be described by a function that depends *only* on the gas temperature T (and *not*, e.g., on the density, pressure, or chemical composition).

In terms of the wave frequency ν , this *Planck Black-Body* function takes the form

$$B_\nu(T) = \frac{2h\nu^3/c^2}{e^{h\nu/kT} - 1}, \quad (4.3)$$

where k is Boltzmann's constant, with value $k = 1.38 \times 10^{-16}$ erg/K = 1.38×10^{-23} Joule/K. For an interval of frequency between ν and $\nu + d\nu$, the quantity $B_\nu d\nu$ gives the emitted energy per unit time per unit area *per unit solid angle*. This means the Planck Black-Body function is fundamentally a measure of *intensity* or *surface brightness*, with B_ν representing the *distribution* of surface brightness over frequency ν , having CGS units erg/cm²/s/ster/Hz (and MKS units W/m²/ster/Hz).

Sometimes it is convenient to instead define this Planck distribution in terms of the brightness distribution in a *wavelength* interval between λ and $\lambda + d\lambda$, $B_\lambda d\lambda$. Requiring that this equals $B_\nu d\nu$, and noting that $\nu = c/\lambda$ implies $|d\nu/d\lambda| = c/\lambda^2$, we can use eqn. (4.3) to obtain

$$B_\lambda(T) = \frac{2hc^2/\lambda^5}{e^{hc/\lambda kT} - 1}. \quad (4.4)$$

4.3 Inverse-temperature dependence of wavelength for peak flux

Figure 4.2 plots the variation of B_λ vs. wavelength λ for various temperatures T . Note that for higher temperature, the level of B_λ is higher at *all* wavelengths, with greatest increases near the peak level.

Moreover, the location of this peak shifts to *shorter* wavelength with *higher* temperature. We can determine this peak wavelength λ_{max} by solving the equation

$$\left[\frac{dB_\lambda}{d\lambda} \right]_{\lambda=\lambda_{max}} \equiv 0. \quad (4.5)$$

Leaving the details as an exercise, the result is

$$\lambda_{max} = \frac{2.9 \times 10^6 \text{ nm K}}{T} = \frac{290 \text{ nm}}{T/10,000 \text{ K}} \approx \frac{500 \text{ nm}}{T/T_\odot}, \quad (4.6)$$

which is known as *Wien's displacement law*.

For example, the last equality uses the fact that the observed wavelength peak in the Sun's spectrum is $\lambda_{max,\odot} \approx 500 \text{ nm}$, very near the middle of the visible spectrum.² We can solve for a Black-Body-peak estimate for the Sun's surface temperature

$$T_\odot = \frac{2.9 \times 10^6 \text{ nm K}}{500 \text{ nm}} = 5800 \text{ K}. \quad (4.7)$$

By similarly measuring the peak wavelength λ_{max} in other stars, we can likewise derive an estimate of their surface temperature by

$$T = T_\odot \frac{\lambda_{max,\odot}}{\lambda_{max}} \approx 5800 \text{ K} \frac{500 \text{ nm}}{\lambda_{max}}. \quad (4.8)$$

4.4 Inferring stellar temperatures from photometric colors

In practice, this is not quite the approach to estimating a star's temperature that is most commonly used in astronomy, in part because with real SEDs, it is relatively difficult to identify accurately the peak wavelength. Moreover in surveying a large number of stars, it requires a lot more effort (and telescope time) to measure the full SED, especially for relatively faint stars. A simpler, more common method is just to measure the stellar *color*.

But rather than using the Red, Green, and Blue (RGB) colors we perceive with our eyes, astronomers typically define a set of standard colors that extend to wavebands beyond just the visible spectrum. The most common example is the Johnson 3-color UBV (Ultraviolet, Blue, Visible) system. The left panel of

² This is not entirely coincidental, since our eyes evolved to use the wavelengths of light for which the solar illumination is brightest.

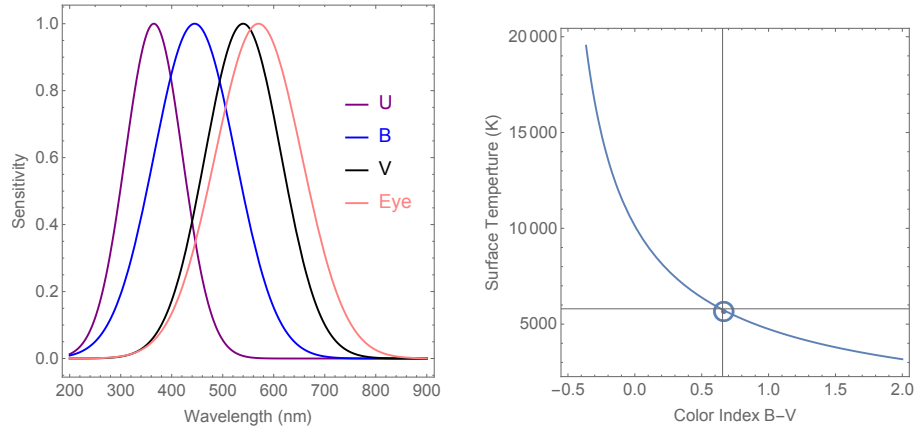


Figure 4.3 *Left:* Comparison of the spectral sensitivity of the human eye with those of the UB filters in the Johnson photometric color system. *Right:* Temperature dependence of the B-V color for a Black-Body emitted spectrum. The circle dot marks the solar values $T_{\odot} \approx 5800$ K and $(B - V)_{\odot} \approx 0.656$.

figure 4.3 compares the wavelength sensitivity of such UB filters to that of the human eye. By passing the star's light through a standard set of filters designed to only let through light for the defined color waveband, the observed apparent brightness in each filter can be used to define a set of color magnitudes, e.g. m_U, m_B , and m_V .

The standard shorthand is simply to denote these color magnitudes just by the capital letter alone, viz. U, B, and V. The *difference* between two color magnitudes, e.g. $B - V \equiv m_B - m_V$, is independent of the stellar distance, but provides a direct diagnostic of the stellar temperature, sometimes called the “color temperature”.

Because a larger magnitude corresponds to a lower brightness, stars with a positive B-V actually are less bright in the blue than in the visible, implying a relatively *low* temperature. On the other hand, a negative B-V means blue is brighter, implying a *high* temperature. The right panel of figure 4.3 shows how the temperature of a Black-Body varies with the B-V color of the emitted Black-Body spectrum.

4.5 Questions and Exercises

Quick Question 1: Two photons have wavelength ratio $\lambda_2/\lambda_1 = 2$.

- What is the ratio of their period P_2/P_1 ?
- What is the ratio of their frequency ν_2/ν_1 ?
- What is the ratio of their energy E_2/E_1 ?

Quick Question 2:

- a. Estimate the temperature of stars with $\lambda_{max} = 100, 300, 1000, \text{ and } 3000 \text{ nm}$. (To simplify the numerics, you may take $T_{\odot} \approx 6000 \text{ K}$.)
- b. Conversely, estimate the peak wavelengths λ_{max} of stars with $T = 2000, 10,000, \text{ and } 60,000 \text{ K}$.
- c. What parts of the EM spectrum (i.e. UV, visible, IR) do each of these lie in?

Quick Question 3:

- a. Assuming the Earth has an average temperature equal to that of typical spring day, i.e. 50°F , compute the peak wavelength of Earth's Black-Body radiation.
- b. What part of the EM spectrum does this lie in?

Exercise 1: Using $B_{\nu}d\nu = B_{\lambda}d\lambda$ and the relationship between frequency ν and wavelength λ , derive eqn. (4.4) from eqn. (4.3).

Exercise 2: Derive eqn. (4.6) from eqn. (4.4) using the definition (4.5).

5 Inferring Stellar Radius from Luminosity and Temperature

We see from figure 4.2 that, in addition to a shift toward shorter peak wavelength λ_{max} , a higher temperature also increases the overall brightness of blackbody emission at *all* wavelengths. This suggests that the total energy emitted over all wavelengths should increase quite sharply with temperature. Leaving the details as an exercise for the reader, let us quantify this expectation by carrying out the necessary spectral integrals to obtain the temperature dependence of the *Bolometric* intensity of a blackbody

$$B(T) \equiv \int_0^\infty B_\lambda(T) d\lambda = \int_0^\infty B_\nu(T) d\nu = \frac{\sigma_{sb} T^4}{\pi}, \quad (5.1)$$

with $\sigma_{sb} = 2\pi^5 k^4 / (15h^3 c^2)$ known as the Stefan-Boltzmann constant, with numerical value $\sigma_{sb} = 5.67 \times 10^{-5} \text{ erg/cm}^2/\text{s/K}^4 = 5.67 \times 10^{-8} \text{ J/m}^2/\text{s/K}^4$.

If we spatially resolve a pure blackbody with surface temperature T , then $B(T)$ represents the Bolometric *surface brightness* we would observe from each part of the visible surface.

5.1 Stefan-Boltzmann law for surface flux from a blackbody

Combining eqns. (3.6) and (5.1), we see that the radiative *flux* at the surface radius R of a blackbody is given by

$$\boxed{F_* \equiv F(R) = \pi B(T) = \sigma_{sb} T^4}, \quad (5.2)$$

which is known as the *Stefan-Boltzman law*.

The Stefan-Boltzmann law is one of the linchpins of stellar astronomy. If we now relate the surface flux to the stellar luminosity L over the surface area $4\pi R^2$, then applying this to the Stefan-Boltzmann law gives

$$\boxed{L = \sigma_{sb} T^4 4\pi R^2}, \quad (5.3)$$

which is often more convenient to scale by associated solar values,

$$\frac{L}{L_\odot} = \left(\frac{T}{T_\odot} \right)^4 \left(\frac{R}{R_\odot} \right)^2. \quad (5.4)$$

We can also use eqn. (5.3) to solve for the stellar radius,

$$R = \sqrt{\frac{L}{4\pi\sigma_{sb}T^4}} = \sqrt{\frac{F(d)}{\sigma_{sb}T^4}} d, \quad (5.5)$$

where the latter equation uses the inverse-square-law to relate the stellar radius to the flux $F(d)$ and distance d , along with the surface temperature T .

For a star with a known distance d , e.g. by a measured parallax, measurement of apparent magnitude gives the flux $F(d)$, while measurement of the peak wavelength λ_{max} or color (e.g. B-V) provides an estimate of the temperature T (see figure 4.3). Applying these in eqn. (5.5), we can thus obtain an estimate of the stellar radius R .

5.2 Questions and Exercises

Quick Question 1: Compute the luminosity L (in units of the solar luminosity L_\odot), absolute magnitude M , and peak wavelength λ_{max} (in nm) for stars with (a) $T = T_\odot$; $R = 10R_\odot$, (b) $T = 10T_\odot$; $R = R_\odot$, and (c) $T = 10T_\odot$; $R = 10R_\odot$. If these stars all have a parallax of $p = 0.001$ arcsec, compute their associated apparent magnitudes m .

Quick Question 2: Suppose a star has a parallax $p = 0.01$ arcsec, peak wavelength $\lambda_{max} = 250$ nm, and apparent magnitude $m = +5$. About what is its:

- Distance d (in pc)?
- Distance modulus $m - M$?
- Absolute magnitude M ?
- Luminosity L (in L_\odot)?
- Surface temperature T (in T_\odot)?
- Radius R (in R_\odot)?
- Angular radius α (in radian and arcsec)?
- Solid angle Ω (in steradian and arcsec^2)?
- Surface brightness relative to that of the Sun, B/B_\odot ?

6 Absorption Lines in Stellar Spectra

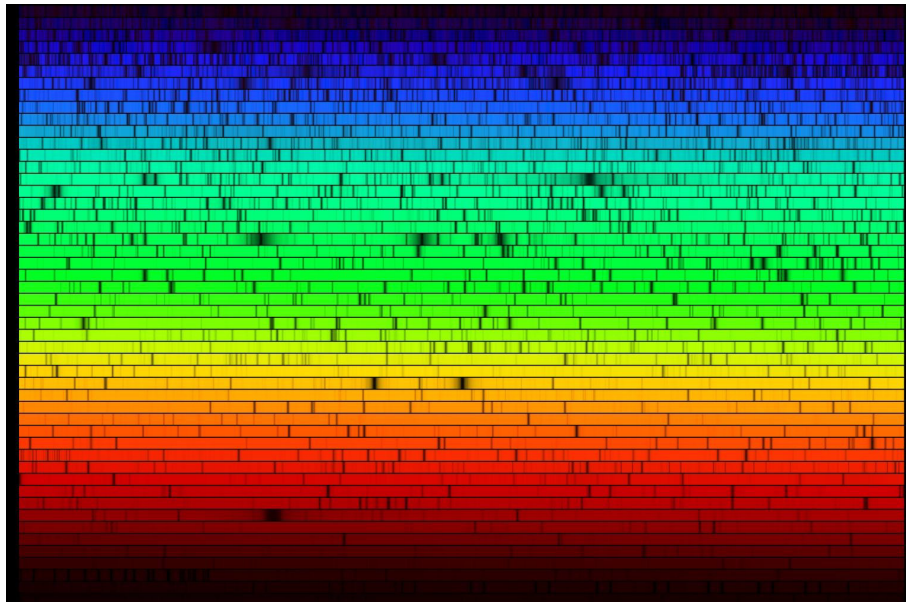


Figure 6.1 The Sun’s spectrum, showing the complex pattern of absorption lines at discrete wavelength or colors. [NOAO/AURA/NSF]

In reality stars are not perfect blackbodies, and so their emitted spectra don’t just depend on temperature, but contain detailed signatures of key physical properties like elemental composition. The energy we see emitted from a stellar surface is generated in the very hot interior and then diffuses outward, following the strong temperature decline to the surface. The atoms and ions that absorb and emit the light don’t do so with perfect efficiency at all wavelengths, which is what is meant by the “black” in “blackbody”. We experience this all the time in our everyday world, which shows that different objects have distinct “color”, meaning they absorb certain wavebands of light, and reflect others. For example, a green leaf reflects some of the “green” parts of the visible spectrum – with wavelengths near $\lambda \approx 5100 \text{ \AA}$ – and absorbs most of the rest.

For atoms in a gas, the ability to absorb, scatter and emit light can likewise

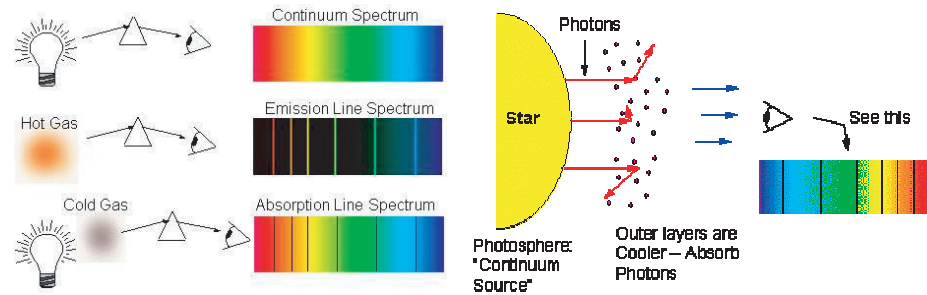


Figure 6.2 Illustration of principals for producing an emission vs. an absorption line spectrum. The left panel shows that an incandescent light passed through a prism generally produces a featureless continuum spectrum, but a cold gas placed in front of this yields an absorption line spectrum. That same gas when heated and seen on its own against a dark background produces the same pattern of lines, but now in *emission* instead of absorption. The right panel shows heuristically how the relatively cool gas in the surface layers of a star leads to an absorption line spectrum from the star.

depend on the wavelength, sometimes quite sharply. Just as the energy of light is quantized into discrete bundles called photons, the energy of electrons orbiting an atomic nucleus have discrete levels, much like the steps in a staircase. Absorption or scattering by the atom is thus much more efficient for those select few photons with an energy that closely matches the energy difference between two of these atomic energy levels.

The evidence for this is quite apparent if we examine carefully the actual spectrum emitted by any star. Although the overall “Spectral Energy Distribution” (SED) discussed above often roughly fits a Planck Black-Body function, careful inspection shows that light is missing or reduced at a number of discrete wavelengths or colors. As illustrated in figure 6.1 for the Sun, when the color spectrum of light is spread out, for example by a prism or diffraction grating, this missing light appears as a complex series of relatively dark “absorption lines”.

Figure 6.2 illustrates how the absorption by relatively cool, low-density atoms in the upper layers of the Sun or a star’s atmosphere can impart this pattern of absorption lines on the continuum, nearly Black-Body spectrum emitted by the denser, hotter layers.

A key point here is that the discrete energies levels associated with atoms of different elements (or, as discussed below, different “ionization stages” of a given element) are quite distinct. As such the associated wavelengths of the absorption lines in a star’s spectrum provide a direct “fingerprint” – perhaps even more akin to a supermarket *bar code* – for the presence of that element in the star’s atmosphere. The code “key” can come from laboratory measurement of the line-spectrum from known samples of atoms and ions, or, as discussed in §A.1, from theoretical models of the atomic energy levels using modern principles of quantum physics.

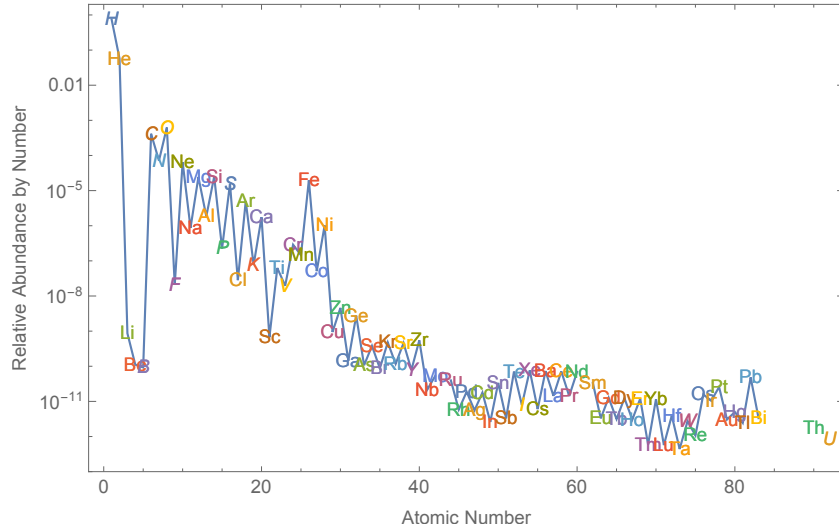


Figure 6.3 Number fractions of elements, plotted on a log scale vs. atomic number, with data points labeled by the symbols for each element.

6.1 Elemental composition of the Sun and stars

With proper physical modeling, the relative strengths of the absorption lines can even provide a quantitative measure of the relative *abundance* of the various elements. A key result is that the composition of the Sun, which is typical of most all stars, is dominated by just the two simplest elements, namely *Hydrogen* (H) and *Helium* (He) – which make up respectively 90.9% and 8.9% of the atoms, with all the other only about 0.2%. Figure 6.3 gives a log plot of these number fractions vs. atomic number.

The corresponding *mass* fractions are $X \approx 0.72$ and $Y \approx 0.26$ for Hydrogen and Helium. All the remaining elements of the periodic table – commonly referred to in astronomy as “metals” – make up just the final two percent of the mass., denoted as a “metallicity” $Z \approx 0.02$. Of these, the most abundant are Oxygen, Carbon, and Iron, with respective mass fractions of 0.009, 0.003, and 0.001.

Like all the planets in our solar system, the Earth formed out of the same material that makes up the Sun (§23). But its relatively weak gravity has allowed a lot of the light elements like Hydrogen and Helium to escape into space, leaving behind the heavier elements that make up our world, and us (§24). Indeed, once the H and He are removed, the *relative* abundances of all these elements are roughly the same on the Earth as in the Sun!

6.2 Stellar spectral type: ionization abundances as temperature diagnostic

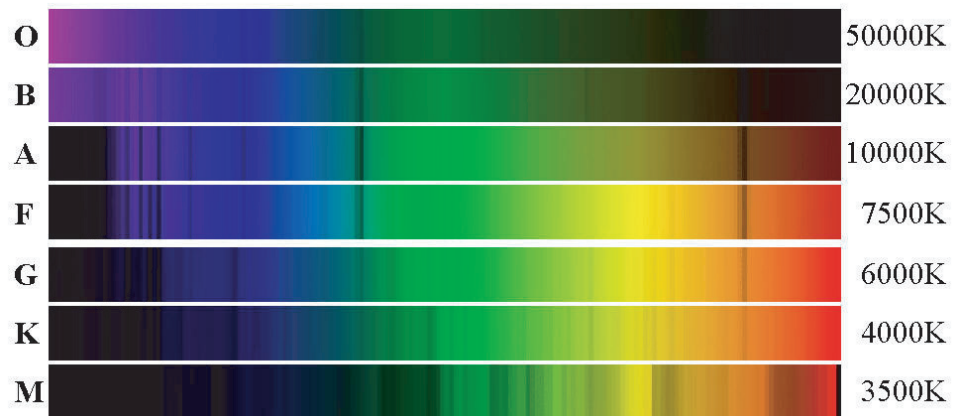


Figure 6.4 Stellar spectra for the full range of spectral types OBAFGKM, corresponding to a range in stellar surface temperature from hot to cool. [NOAO/AURA/NSF]

Another key factor in the observed stellar spectra is that the atomic elements present are generally not electrically neutral, but typically have had one or more electrons stripped – ionized – by thermal collisions with characteristic energies set by the temperature. As such, the observed degree of ionization depends on the temperature near the visible stellar surface. Figure 6.4 compares the spectra of stars of different surface temperature, showing that this leads to gradual changes and shifts in the detailed pattern of absorption lines from the various ionizations stages of the various elements. The letters “OBAFGKM” represent various categories, known as spectral class or “spectral type”, assigned to stars with different spectral patterns. It turns out that type O is the hottest, with temperatures about 50,000 K, while M is the coolest¹ with temperatures of about 3500 K. The sequence is often remembered through the mnemonic² “Oh, Be A Fine Gal/Guy Kiss Me”. In keeping with its status as a kind of average star, the Sun has spectral type G, just a bit cooler than type F in the middle of the sequence.

In addition to the spectral classes OBAFGKM that depend on surface temperature T , spectra can also be organized in terms of *luminosity* classes, convention-

¹ In recent years, it has become possible to detect even cooler “Brown dwarf” stars, with spectral classes LTY, extending down to temperatures as low as 1000 K. Brown dwarf stars have too low a mass ($< 0.08M_{\odot}$) to force hydrogen fusion in their interior (see §16.3). They represent a link to gas giant planets like Jupiter (for which $M_J \approx 0.001M_{\odot}$).

² A student in one of my exams once offered an alternative mnemonic: “Oh Boy, Another F’s Gonna Kill Me”.

ally denoted though Roman numerals I for the biggest, brightest “supergiant” stars, to V for smaller, dimmer “dwarf” stars; in between, there are luminosity classes II (bright giants), III (giants), and IV (sub-giants).

In this two-parameter scheme, the Sun is classified as a G2V star.

Finally, in addition to giving information on the temperature, chemical composition, and other conditions of a star’s atmosphere, these absorption lines provide convenient “markers” in the star’s spectrum. As discussed in §9.2, this makes it possible to track small changes in the wavelength of lines that arise from the so-called Doppler effect as a star moves toward or away from us.

In summary, the appearance of absorption lines in stellar spectra provides a real treasure trove of clues to the physical properties of stars.

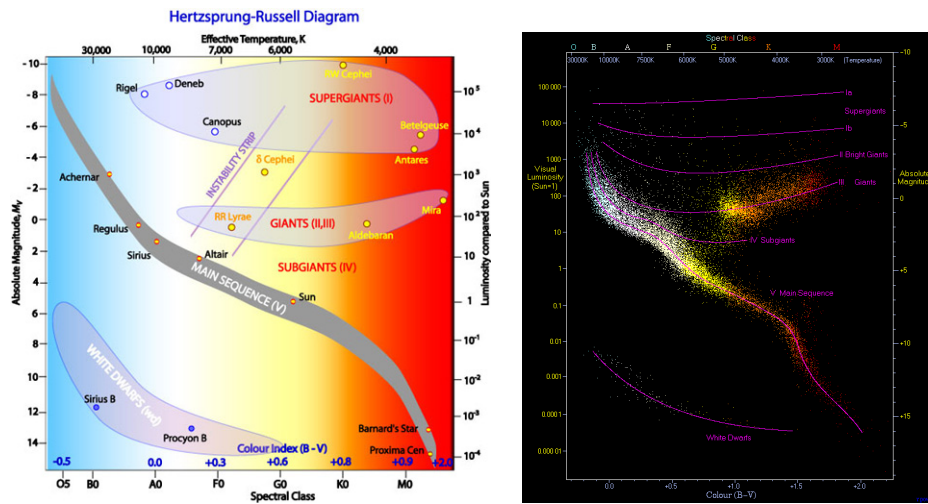


Figure 6.5 *Left:* Hertzsprung-Russell (H-R) diagram relating star’s absolute magnitude (or log luminosity) vs. surface temperature, as characterized by the spectral type or color, with hotter bluer stars on the left, and cooler redder stars on the right. The main sequence (MS) represents stars burning Hydrogen into Helium in their core, whereas the giants and supergiants are stars that have evolved away from the MS after exhausting Hydrogen in their cores. The White Dwarf stars are dying remnants of solar-type stars. *Right:* Observed H-R diagram for stars in the solar neighborhood. The points include 22,000 stars from the Hipparcos Catalogue together with 1000 low-luminosity stars (red and white dwarfs) from the Gliese Catalogue of Nearby Stars.

6.3 Hertzsprung-Russell (H-R) diagram

A key diagnostic of stellar populations comes from the *Hertzsprung-Russell* (H-R) diagram, illustrated by the left panel of figure 6.5. Observationally, it relates (absolute) magnitude (or luminosity class) on the y-axis, to color or spectral

type on the x-axis; physically, it relates luminosity to temperature. For stars in the solar neighborhood with parallaxes measured by the *Hipparchus* astrometry satellite, one can readily use the associated distance to convert observed apparent magnitudes to absolute magnitudes and luminosities. The right panel of figure 6.5 shows the H-R diagram for these stars, plotting their known luminosities vs. their colors or spectral types, with the horizontal lines showing the luminosity classes³.

The extended band of stars running from the upper left to lower right is known as the *main sequence*, representing “dwarf” stars of luminosity class V. The reason there are so many stars in this main-sequence band is that it represents the long-lived phase when stars are stably burning Hydrogen into Helium in their cores (§18).

The medium horizontal band above the main sequence represents “giant stars” of luminosity class III. They are typically stars that have exhausted hydrogen in their core, and are now getting energy from a combination of hydrogen burning in a shell around the core, and burning Helium into Carbon in the cores themselves (§19).

The relative lack here of still more luminous supergiant stars of luminosity class I stems from both the relative rarity of stars with sufficiently high mass to become this luminous, coupled with the fact that such luminous stars only live for a very short time (§8.4). As such, there are only a few such massive, luminous stars in the solar neighborhood. Studying them requires broader surveys extending to larger distances that encompass a greater fraction of our galaxy.

The stars in the band below the main sequence are called *white dwarfs*; they represent the slowly cooling remnant cores of low-mass stars like the Sun (§19.4).

This association between position on the H-R diagram, and stellar parameters and evolutionary status, represents a key link between the observable properties of light emitted from the stellar surface and the physical properties associated with the stellar interior. Understanding this link through examination of stellar structure and evolution will constitute the major thrust of our studies of stellar interiors in part II of these notes.

But before we can do that, we need to consider ways that we can empirically determine the two key parameters differentiating the various kinds of stars on this H-R diagram, namely *mass* and *age*.

6.4 Questions and Exercises

Quick Question 1: On the H-R diagram, where do we find stars that are: a.) Hot and luminous? b.) Cool and luminous? c.) Cool and Dim? d.) Hot and Dim?

Which of these are known as: 1.) White Dwarfs? 2.) Red Giants? 3.) Blue supergiants? 4.) Red dwarfs?

³ The more recent **GAIA** satellite has provided an even more extensive H-R diagram representing more than 4 million stars within 5000 pc. See <https://sci.esa.int/web/gaia/-/60198-gaia-hertzsprung-russell-diagram>.

7 Surface Gravity and Escape/Orbital Speed

So far we've been able to find ways to estimate the first five stellar parameters on our list – distance, luminosity, temperature, radius, and elemental composition. Moreover, we've done this with just a few, relatively simple measurements – parallax, apparent magnitude, color, and spectral line patterns. But along the way we've had to learn to exploit some key geometric principles and physical laws – angular-size/parallax, inverse-square law, and Planck's, Wien's and the Stefan-Boltzman laws of blackbody radiation.

So what of the next item on the list, namely stellar mass? Mass is clearly a physically important parameter for a star, since for example it will help determine the strength of the gravity that tries to pull the star's matter together. To lay the groundwork for discussing one basic way we can determine mass (from orbits of stars in stellar binaries), let's first review Newton's law of gravitation and show how this sets such key quantities like the surface gravity, and the speeds required for material to escape or orbit the star.

7.1 Newton's law of gravitation and stellar surface gravity

On Earth, an object of mass m has a weight given by

$$F_{grav} = mg_e, \quad (7.1)$$

where the acceleration of gravity on Earth is $g_e = 980 \text{ cm/s}^2 = 9.8 \text{ m/s}^2$. But this comes from Newton's law of gravity, which states that for two point masses m and M separated by a distance r , the attractive gravitational force between them is given by

$$F_{grav} = \frac{GMm}{r^2}, \quad (7.2)$$

where Newton's constant of gravity is $G = 6.7 \times 10^{-8} \text{ cm}^3/\text{g/s}^2$. Remarkably, when applied to spherical bodies of mass M and finite radius R , the same formula works for all distances $r \geq R$ at or outside the surface!¹ Thus, we see that the

¹ Even more remarkably, even if we are *inside* the radius, $r < R$, then we can still use Newton's law if we just count that part of the total mass that is *inside* r , i.e. M_r , and completely ignore all the mass that is above r .

acceleration of gravity at the surface of the Earth is just given by the mass and radius of the Earth through

$$g_e = \frac{GM_e}{R_e^2}. \quad (7.3)$$

Similarly for stars, the surface gravity is given by the stellar mass M and radius R . In the case of the Sun, this gives $g_\odot = 2.6 \times 10^4 \text{ cm/s}^2 \approx 27 g_e$. Thus, if you could stand on the surface of the Sun, your “weight” would be about 27 times what it is on Earth.

For other stars, gravities can vary over a quite wide range, largely because of the wide range in size. For example, when the Sun gets near the end of its life about 5 billion years from now, it will swell up to more than 100 times its current radius, becoming what’s known as a “Red Giant” (§19). Stars we see now that happen to be in this Red Giant phase thus tend to have quite low gravity, about a fraction 1/10,000 that of the Sun.

Largely because of this very low gravity, much of the outer envelope of such Red Giant stars will actually be lost to space (forming, as we shall see, quite beautiful nebulae; see §19 and figure 20.5.) When this happens to the Sun, what’s left behind will be just the hot stellar core, a so-called “white dwarf”, with about 2/3 the mass of the current Sun, but with a radius only about that of the Earth, i.e. $R \approx R_e \approx 7 \times 10^3 \text{ km} \approx 0.01 R_\odot$. The surface gravities of white dwarfs are thus typically 10,000 times *higher* than the current Sun (§19.4).

For “neutron stars”, which are the remnants of stars a bit more massive than the Sun, the radius is just about 10 km, more than another factor 500 smaller than white dwarfs (§20.3). This implies surface gravities another 5-6 orders of magnitude higher than even white dwarfs. (Imagine what you’d weigh then on the surface of a neutron star!)

Since stellar gravities vary over such a large range, it is customary to quote them in terms of the log of the gravity, $\log g$, using CGS units. We thus have gravities ranging from $\log g \approx 0$ for Red Giants, to $\log g \approx 4$ for normal stars like the Sun, to $\log g \approx 8$ for white dwarfs, to $\log g \approx 13$ for neutron stars. Since the Earth’s gravity has $\log g_e \approx 3$, the difference of $\log g$ from 3 is the number of order of magnitudes more/less that you’d weigh on that surface. For example, for neutron stars the difference from Earth is 10, implying you’d weigh 10^{10} , or ten billion times more on a neutron star! On the other hand, on a Red Giant, your weight would be about 1000 times *less* than on Earth.

7.2 Surface escape speed V_{esc}

Another measure of the strength of a gravitational field is through the surface escape speed,

$$V_{esc} = \sqrt{\frac{2GM}{R}}. \quad (7.4)$$

A object of mass m launched with this speed has a kinetic energy $mV_{esc}^2/2 = GMm/R$. This just equals the work needed to lift that object from the surface radius R to escape at a large radius $r \rightarrow \infty$,

$$W = \int_R^\infty \frac{GMm}{r^2} dr = \frac{GMm}{R}. \quad (7.5)$$

Thus if one could throw a ball (or launch a rocket!) with this speed outward from a body's surface radius R , then² by conservation of total energy, that object would reach an arbitrarily large distance from the star, with however a vanishingly small final speed.

For the earth, the escape speed is about 25,000 mph, or 11.2 km/s. By comparison, for the moon, it is just 2.4 km/s, which is one reason the Apollo astronauts could use a much smaller rocket to get back from the moon, than they used to get there in the first place. However, escaping from the surface of the Sun (and most any star), is *much* harder, requiring an escape speed of 618 km/s.

7.3 Speed for circular orbit

Let us next compare this escape speed with the speed needed for an object to maintain a circular orbit at some radius r from the center a gravitating body of mass M . For an orbiting body of mass m , we require that the gravitational force be balanced by the centrifugal force from moving along the circle of radius r ,

$$\frac{GMm}{r^2} = \frac{mV_{orb}^2}{r}, \quad (7.6)$$

which solves to

$$V_{orb}(r) = \sqrt{\frac{GM}{r}}. \quad (7.7)$$

Note in particular that the orbital speed very near the stellar surface, $r \approx R$, is given by $V_{orb}(R) = V_{esc}/\sqrt{2}$. Thus the speed of satellites in low-earth-orbit (LEO) is about 17,700 mph, or 7.9 km/s.

Of course, orbits can also be maintained at any radius above the surface radius, $r > R$, and eqn. (7.7) shows that in this case, the speed needed declines as $1/\sqrt{r}$. Thus, for example, the orbital speed of the earth around the Sun is about 30 km/s, a factor of $\sqrt{R_\odot/au} = \sqrt{1/215} = 0.0046$ smaller than the orbital speed near the Sun's surface, $V_{orb,\odot} = 434$ km/s.

7.4 Virial Theorem for bound orbits

If we define the gravitational energy to be zero far from a star, then for an object of mass m at a radius r from a star of mass M , we can write the gravitational

² neglecting forces other than gravity, like the drag from an atmosphere

binding energy U as the *negative* of the escape energy,

$$U(r) = -\frac{GMm}{r}. \quad (7.8)$$

If this same object is in orbit at this radius r , then the kinetic energy of the orbit is

$$T(r) = \frac{mV_{orb}^2}{2} = +\frac{GMm}{2r} = -\frac{U(r)}{2}, \quad (7.9)$$

where the second equation uses eqn. (7.7) for the orbital speed $V_{orb}(r)$. We can then write the *total* energy as

$$E(r) \equiv T(r) + U(r) = -T(r) = \frac{U(r)}{2}. \quad (7.10)$$

This fact that the total energy E just equals *half* the gravitational binding energy U is an example of what is known as the *Virial Theorem*. It is applicable broadly to most any stably bound gravitational system. For example, if we recognize that the thermal energy inside a star as a kind of kinetic energy, it even applies to stars, in which the internal gas pressure balances the star's own self gravity. This is discussed further in §8.2 and the part II notes on stellar structure.

7.5 Questions and Exercises

Quick Question 1: In CGS units, the Sun has $\log g_{\odot} \approx 4.44$. Compute the $\log g$ for stars with:

- a. $M = 10M_{\odot}$ and $R = 10R_{\odot}$
- b. $M = 1M_{\odot}$ and $R = 100R_{\odot}$
- c. $M = 1M_{\odot}$ and $R = 0.01R_{\odot}$

Quick Question 2:

The Sun has an escape speed of $V_{e\odot} = 618 \text{ km/s}$. Compute the escape speed V_e of the stars in parts a-c of QQ1.

Quick Question 3:

The earth has an orbital speed of $V_e = 2\pi \text{ au/yr} = 30 \text{ km/s}$. Compute the orbital speed V_{orb} (in km/s) of a body at the following distances from the stars with the quoted masses:

- a. $M = 10M_{\odot}$ and $d = 10 \text{ au}$.
- b. $M = 1M_{\odot}$ and $d = 100 \text{ au}$.
- c. $M = 1M_{\odot}$ and $d = 0.01 \text{ au}$.

Exercise 1:

a. During a solar eclipse, the moon just barely covers the visible disk of the Sun. What does this tell you about the relative angular size of the Sun and moon?

b. Given that the moon is at a distance of 0.0024 au, what then is the ratio of the *physical* size of the moon vs. Sun?

c. Compared to earth, the Sun and moon have gravities of respectively $27g_e$ and $g_e/6$. Using this and your answer above, what is the ratio of the *mass* of the moon to that of the Sun?

d. Using the above, plus known values for Newton's constant G , earth's gravity $g_e = 9.8 \text{ m/s}^2$, and the solar radius $R_\odot = 700,000 \text{ km}$, compute the masses of the Sun and moon in kg.

Exercise 2:

a. What is the ratio of the *energy* needed to escape the moon vs. the earth? What's the ratio for the Sun vs. the earth?

b. What is the escape speed (in km/s) from a star with: (1) $M = 10M_\odot$ and $R = 10R_\odot$; (2) $M = 1M_\odot$ and $R = 100R_\odot$; (3) $M = 1M_\odot$ and $R = 0.01R_\odot$?

c. To what radius (in km) would you have to shrink the Sun to make its escape speed equal to the speed of light c ?

Exercise 3:

a. What is the ratio of the *energy* needed to escape the the earth vs. that needed to reach LEO?

b. What is the orbital speed (in km/s) of a planet that orbits at a distance a from a star with mass M , given: (1) $M = 10M_\odot$ and $a = 10 \text{ au}$; (2) $M = 1M_\odot$ and $a = 100 \text{ au}$; (3) $M = 1M_\odot$ and $a = 0.01 \text{ au}$?

8 Stellar Ages and Lifetimes

In our list of basic stellar properties, let us next consider stellar age. Just how old are stars like the Sun? What provides the energy that keeps them shining? And what will happen to them as they exhaust various available energy sources?

8.1 Shortness of chemical burning timescale for Sun and stars

When 19th century scientists pondered the possible energy sources for the Sun, some first considered whether this could come from the kind of chemical reactions (e.g., from fossil fuels like coal, oil, natural gas, etc.) that power human activities on Earth. But such chemical reactions involve transitions of electrons among various bound states of atoms, and, as discussed below (§A.1) for the Bohr model of the Hydrogen, the scale of energy release in such transitions is limited to something on the order of an electron volt (eV). In contrast, the rest-mass energy of the protons and neutrons that make up the mass is about 1 GeV, or 10^9 times higher. With the associated mass-energy efficiency of $\epsilon \sim 10^{-9}$, we can readily estimate a timescale for maintaining the solar luminosity from chemical reactions,

$$t_{chem} = \epsilon \frac{M_{\odot} c^2}{L_{\odot}} = \epsilon 4.5 \times 10^{20} \text{ s} = \epsilon 1.5 \times 10^{13} \text{ yr} \approx 15,000 \text{ yr} . \quad (8.1)$$

Even in the 19th century, it was clear, e.g. from geological processes like erosion, that the Earth – and so presumably also the Sun – had to be much older than this.

8.2 Kelvin-Helmholtz timescale for gravitational contraction

So let us consider whether, instead of chemical reactions, gravitational contraction might provide the energy source to power the Sun and other stars. As a star undergoes a contraction in radius, its gravitational binding becomes stronger, with a deeper gravitational potential energy, yielding an energy release set by the negative of the change in gravitational potential ($-dU > 0$). If the contraction is gradual enough that the star roughly maintains dynamical equilibrium,

then just half of the gravitational energy released goes into heating up the star¹, leaving the other half available to power the radiative luminosity, $L = -\frac{1}{2}dU/dt$. For a star of observed luminosity L and present-day gravitational binding energy U , we can thus define a characteristic gravitational contraction lifetime,

$$t_{grav} = -\frac{1}{2} \frac{U}{L} \equiv t_{KH} \quad (8.2)$$

where the subscript “KH” refers to Kelvin and Helmholtz, the names of the two scientists credited with first identifying this as an important timescale. To estimate a value for the gravitational binding energy, let us consider the example for the Sun under the somewhat artificial assumption that it has a uniform, constant density, given by its mass over volume, $\rho = M_\odot/(4\pi R_\odot^3/3)$. Since the gravity at any radius r depends only on the mass $m = \rho 4\pi r^3/3$ inside that radius, the total gravitational binding energy of the Sun is given by integrating the associated local gravitational potential $-Gm/r$ over all differential mass shells dm ,

$$-U = \int_0^{M_\odot} \frac{Gm}{r} dm = \frac{16\pi^2}{3} G\rho^2 \int_0^R r^4 dr = \frac{3}{5} \frac{GM_\odot^2}{R_\odot}, \quad (8.3)$$

Applying this in eqn. (8.2), we find for the “Kelvin-Helmholtz” time of the Sun,

$$t_{KH} \approx \frac{3}{10} \frac{GM_\odot^2}{R_\odot L_\odot} \approx 30 \text{ Myr}. \quad (8.4)$$

Although substantially longer than the chemical burning timescale (8.1), this is still much shorter than the geologically inferred minimum age of the Earth, which is several *Billion* years.

8.3 Nuclear burning timescale

We now realize, of course, that the ages and lifetimes of stars like the Sun are set by a much longer *nuclear burning* timescale. When four hydrogen nuclei are fused into a helium nucleus, the helium mass is about 0.7% *lower* than the original four hydrogen. For nuclear fusion the above-defined mass-energy burning efficiency is thus now $\epsilon_{nuc} \approx 0.007$. But in a typical main sequence star, only some core fraction $f \approx 1/10$ of the stellar mass is hot enough to allow Hydrogen fusion. Applying this we thus find for the nuclear burning timescale

$$t_{nuc} = \epsilon_{nuc} f \frac{Mc^2}{L} \approx 10 \text{ Gyr} \frac{M/M_\odot}{L/L_\odot}, \quad (8.5)$$

where $\text{Gyr} \equiv 10^9 \text{ yr}$, i.e., a billion years, or a “Giga-year”.

¹ This is another example of the Virial theorem for gravitationally bound systems, as discussed in 7.4.

We thus see that the Sun can live for about 10 Gyr by burning Hydrogen into Helium in its core. It's present age of 4.6 Gyr² thus puts it roughly half way through this Hydrogen-burning phase, with about 5.4 Gyr to go before it runs out of H in its core.

8.4 Age of stellar clusters from main-sequence turnoff point

As discussed below (see §10.4 and eqn. 10.11), observations of stellar binary systems indicate that the luminosities of main-sequence stars scale with a high power of the stellar mass – roughly $L \sim M^3$. In the present context, this implies that high-mass stars should have much shorter lifetimes than low-mass stars.

If we make the reasonable assumption that the same fixed fraction ($f \approx 0.1$) of the total hydrogen mass of any star is available for nuclear burning into helium in its stellar core, then the fuel available scales with the mass, while the burning rate depends on the luminosity. Normalized to the Sun, the main-sequence lifetime thus scales as

$$t_{ms} = t_{ms,\odot} \frac{M/M_\odot}{L/L_\odot} \approx 10 \text{ Gyr} \left(\frac{M_\odot}{M} \right)^2. \quad (8.6)$$

The most massive stars, of order $100 M_\odot$, and thus with luminosities of order $10^6 L_\odot$, have main-sequence lifetimes of only about 1 Myr, much shorter the multi-Gyr timescale for solar-mass stars.

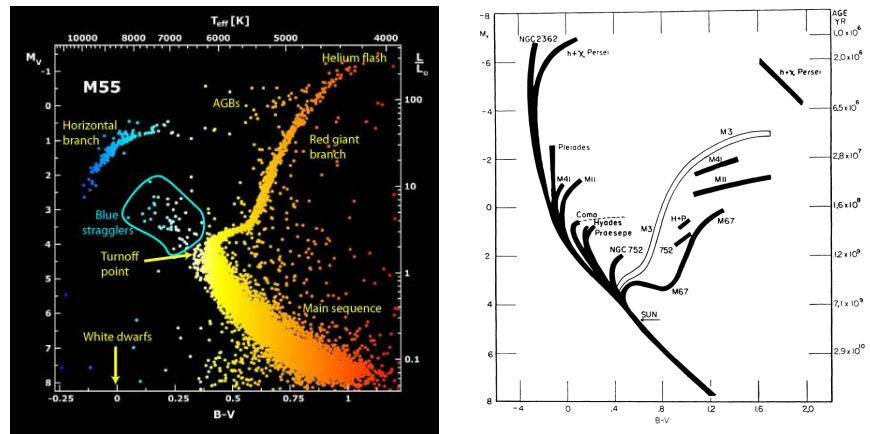


Figure 8.1 Left: H-R diagram for globular cluster M55, showing how stars on the upper main sequence have evolved to lower temperature giant stars. Right: Schematic H-R diagram for clusters, showing the systematic peeling off of the main sequence with increasing cluster age.

² As inferred, e.g., from radioactive dating of the oldest meteorites.

This strong scaling of lifetime with mass can be vividly illustrated by plotting the H-R diagram of stellar clusters. The H-R diagram plotted in figure 6.5 is for volume-limited sample near the Sun, consisting of stars of a wide range of ages, distances, and perhaps even chemical composition. But stars often appear in clusters, all roughly at the same distance, and, since they likely formed over a relatively short time span out of the same interstellar cloud, they all have roughly the same age and chemical composition. Using eqn. (8.6) together with the $L \sim M^3$ relation, the age of a stellar cluster can be inferred from its H-R diagram simply by measuring the luminosity L_{to} of stars at the “turn-off” point from the main sequence,

$$t_{cluster} \approx 10 \text{ Gyr} \left(\frac{L_{\odot}}{L_{to}} \right)^{2/3}. \quad (8.7)$$

The left panel of figure 8.1 plots an actual H-R diagram for the globular cluster M55. Note that stars to the upper left of the main sequence have evolved to a vertical branch of cooler stars extending up to the Red Giants³. This reflects the fact that more luminous stars exhaust their hydrogen fuel sooner than dimmer stars, as shown by the inverse luminosity scaling of the nuclear burning timescale in eqn. (8.5). The right panel illustrates schematically the H-R diagrams for various types of stellar clusters, showing how the turnoff point from the main sequence is an indicator of the cluster age. Observed cluster H-R diagrams like this thus provide a direct diagnostic of the formation and evolution of stars with various masses and luminosities.

8.5 Questions and Exercises

Quick Question 1: What are the luminosities (in L_{\odot}) and the expected main sequence lifetimes (in Myr) of stars with masses: a. $10 M_{\odot}$? b. $0.1 M_{\odot}$? c. $100 M_{\odot}$?

Quick Question 2: Suppose you observe a cluster with a main-sequence turnoff point at a luminosity of $100L_{\odot}$. What is the cluster’s age, in Myr. What about for a cluster with a turnoff at a luminosity of $10,000L_{\odot}$?

Exercise 1: A cluster has a main-sequence turnoff at a spectral type $G2$, corresponding to stars of apparent magnitude $m = +10$.

- (a) About what is the luminosity, in L_{\odot} , of the stars at the turnoff point?
- (b) About what is the age (in Gyr) of the cluster?
- (c) About what is the distance (in pc) of the cluster?

³ Stars just above this main sequence turn-off are dubbed “blue stragglers”. They are stars whose close binary companion became a Red Giant with a such big radius that mass from its envelope spilled over onto it. This rejuvenated the mass gainer, making it again a hot, luminous blue star.

Exercise 2: Confirm the integration result in eqn. (8.3).

9 Inferring Stellar Space Velocities

The next section (§10) will use the inferred orbits of stars in *binary* star systems to directly determine stellar masses. But first, as a basis for interpreting observations of such systems in terms of the orbital velocity of the component stars, let us review the astrometric and spectrometric techniques used to measure the motion of stars through space.

9.1 Transverse speed from proper motion observations

In addition to such periodic motion from binary orbits, stars generally also exhibit some systematic motion relative to the Sun, generally with components both transverse (i.e. perpendicular) to and along (parallel to) the observed line of sight. For nearby stars, the perpendicular movement, called “proper motion”, can be observed as a drift in the apparent position in the star relative to the more fixed pattern of more distant, background stars. Even though the associated physical velocities can be quite large, e.g. $V_t \approx 10 - 100$ km/s, the distances to stars is so large that proper motions of stars – measured as an angular drift per unit time, and generally denoted with the symbol μ – are generally no bigger than about $\mu \approx 1$ arcsec/year. But because this is a systematic drift, the longer the star is monitored, the smaller the proper motion that can be detected, down to about $\mu \approx 1$ arcsec/century or less for the most well-observed stars.

Figure 9.1 illustrates the proper motion for Barnard’s star, which has the highest μ value of any star in the sky. It is so high in fact, that its proper motion can even be followed with a backyard telescope, as was done for this figure. This star is actually tracking along the nearly South-to-North path labeled as the “*Hipparcos*¹ mean” in the figure. The apparent, nearly East-West (EW) wobble is due to the Earth’s own motion around the Sun, and indeed provides a measure of the star’s parallax, and thus its distance. Referring to the arcsec marker in the lower right, we can estimate the full amplitude of the wobble at a bit more than an arcsec, meaning the parallax² is $p \approx 0.55$ arcsec, implying a distance

¹ *Hipparcos* is an orbiting satellite that, because of the absence of the atmospheric blurring, can make very precise “astrometric” measurements of stellar positions, at precisions approaching a milli-arcsec.

² given by half the full amplitude, since parallax assumes a 1 au baseline that is half the full diameter of earth’s orbit

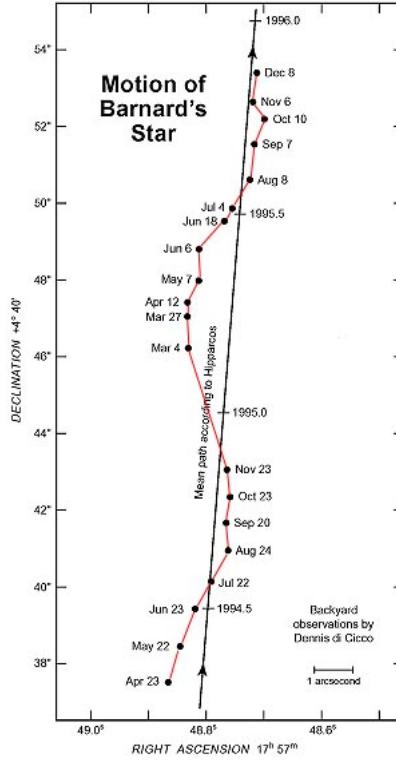


Figure 9.1 Proper motion of Barnard's star. The star is actually tracking along the path labeled as the mean from the *Hipparcos* astrometric satellite. The apparent wobble is due to the parallax from the Earth's own motion around the Sun. Referring to lower right label showing one arcsec, we can estimate the full amplitude of the parallax wobble as about 1.1 arcsec; but since this reflects a baseline of 2 AU from the earth's orbital diameter, the (one-AU) parallax angle is half this, or $p = 0.55$ arcsec, implying a distance of $d = 1/p \approx 1.8$ pc.

of $d \approx 1.8$ pc. By comparison, the roughly South-to-North proper motion has a value $\mu \approx 10$ arcsec/yr.

In general, with a known parallax p in arcsec, and known proper motion μ in arcsec/yr, we can derive the associated transverse velocity V_t across our line of sight,

$$V_t = \frac{\mu}{p} \text{ au/yr} = 4.7 \frac{\mu}{p} \text{ km/s}, \quad (9.1)$$

where the last equality uses the fact that the Earth's orbital speed $V_E = 2\pi \text{ au/yr} = 30 \text{ km/s}$. For Barnard's star this works out to give $V_t \approx 90 \text{ km/s}$, or about 3 times the earth's orbital speed around the Sun. This among the fastest transverse speeds inferred among the nearby stars.

9.2 Radial velocity from Doppler shift

We've seen how we can directly measure the transverse motion of relatively nearby, fast-moving stars in terms of their proper motion. But how might we measure the *radial* velocity component *along* our line of sight? The answer is: via the “Doppler effect”, wherein such radial motion leads to an observed shift in the wavelength of the light.

To see how this effect comes about, we need only consider some regular signal with period P_o being emitted from an object moving at a speed V_r toward ($V_r < 0$) or away ($V_r > 0$) from us. Let the signal travel at a speed V_s , where $V_s = c$ for a light wave, but might equally as well be speed of sound if we were to use that as an example. For clarity of language, let us assume the object is moving away, with $V_r > 0$. Then after any given pulse of the signal is emitted, the object moves a distance $V_r P_o$ before emitting the next pulse. Since the pulse still travels at the same speed, this implies it takes the second pulse an extra time

$$\Delta P = \frac{V_r P_o}{V_s} \quad (9.2)$$

to reach us. Thus the period we observe is longer, $P' = P_o + \Delta P$.

For a wave, the wavelength is given by $\lambda = P V_s$, implying then an associated stretch in the observed wavelength

$$\lambda' = P' V_s = (P_o + \Delta P) V_s = (V_s + V_r) P_o = \lambda_o + V_r P_o. \quad (9.3)$$

where $\lambda_o = P_o V_s$ is the rest wavelength. The associated relative stretch in wavelength is thus just

$$\frac{\Delta \lambda}{\lambda_o} \equiv \frac{\lambda' - \lambda_o}{\lambda_o} = \frac{V_r}{V_s}. \quad (9.4)$$

For sound waves, this formula works in principle as long as $V_r > -V_s$. But if an object moves *toward* us faster than sound ($V_r < -V_s$), then it can basically “overrun” the signal. This leads to strongly compressed sound waves, called “shock waves”, which are the basic origin of the sonic boom from a supersonic jet. For some nice animations of this, see

<http://www.lon-capa.org/~mmp/applist/doppler/d.htm>

A common example of the Doppler effect in sound is the shift in pitch we hear as the object moves past us. Consider the noise from a car on a highway, for which the “vvvvrrrrrooomm” sound stems from just this shift in pitch from the car engine noise. Figure 9.2 illustrates this for a racing car.

In the case of light $V_s = c$, and so we can define the Doppler shift of light as

$$\boxed{\frac{\Delta \lambda}{\lambda_o} = \frac{V_r}{c} \quad ; \quad |V_r| \ll c.} \quad (9.5)$$

This assumes the non-relativistic case that $|V_r| \ll c$, which applies well to most all stellar motions. Straightforward observations of the associated wavelengths of

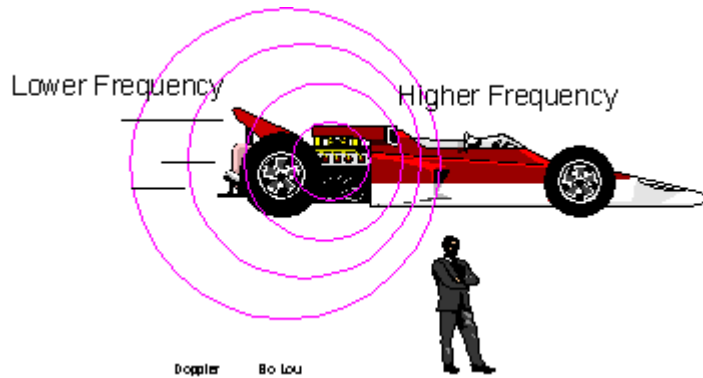


Figure 9.2 Illustration of the Doppler shift of the sound from a racing car.

spectral lines in the star's spectrum relative to their rest (laboratory measured) wavelengths thus gives a direct measurement of the star's motion toward or away from the observer.

For our above example of Barnard's star, observations of the stellar spectrum show a constant *blueshift* of $\Delta\lambda/\lambda = -3.7 \times 10^{-4}$, implying the star is moving *toward* us, with a speed $V_r = zc = -111$ km/s. This allows us to derive the overall *space velocity*,

$$V = \sqrt{V_r^2 + V_t^2}. \quad (9.6)$$

For Barnard's star, this gives $V = 143$ km/s, which again is one of the highest space velocities among nearby stars. Mapping the space motion of nearby stars relative to the Sun provides some initial clues about the kinematics of stars in our local region of the Milky Way galaxy.

9.3 Questions and Exercises

Quick Question 1: A star with parallax $p = 0.02$ arcsec is observed over 10 years to have shifted by 2 arcsec from its proper motion. Compute the star's tangential space velocity V_t , in km/s.

Quick Question 2: For the star in QQ#1, a line with rest wavelength $\lambda_o = 600.00$ nm is observed to be at a wavelength $\lambda = 600.09$ nm.

- Is the star moving toward us or away from us?
- What is the star's Doppler shift z ?
- What is the star's radial velocity V_r , in km/s?
- What is the star's total *space velocity* V_{tot} , in km/s?

10 Using Binary Systems to Determine Masses and Radii

Let us next consider how we can infer the masses of stars, namely through the study of stellar *binary systems*.

It turns out, in fact, that stellar binary (and even triple and quadruple) systems are quite common, so much so that astronomers sometimes joke that “three out of every two stars is (in) a binary”. The joke here works because often two stars in a binary are so close together on the sky that we can’t actually resolve one star from another, and so we sometimes mistake the light source as coming from a single star, when in fact it actually comes from two (or even more). But even in such close binaries, we can often still tell there are two stars by carefully studying the observed spectrum, and in this case, we call the system a “spectroscopic binary” (see the next subsection, and figure 10.2).

But for now, let’s first focus on the simpler example of “visual binaries”, a.k.a. “astrometric” binaries (see figure 10.1), since their detection typically requires precise astrometric measurements of small variations of their positions on the sky over time.

10.1 Visual binaries

In visual binaries, monitoring of the stellar positions over years and even decades reveals that the two stars are actually moving around each other, much as the Earth moves around the Sun. Figure 10.1 illustrates the principles behind visual binaries. The time it takes the stars to go around a full cycle, called the orbital period, can then be measured quite directly. Then if we can convert the apparent angular separation into a physical distance apart – e.g. if we know the distance to the system independently through a measured annual parallax for the stars in the system – then we can use Kepler’s 3rd law of orbital motion (as generalized by Newton) to measure the total mass of the two stars.

It’s actually quite easy to derive the full formula in the simple case of circular orbits that lie in a plane perpendicular to our line of sight. For stars of mass M_1 and M_2 separated by a physical distance a , Newton’s law of gravity gives the attractive force each star exerts on the other,

$$F_g = \frac{GM_1M_2}{a^2}. \quad (10.1)$$

Binary Star Orbit

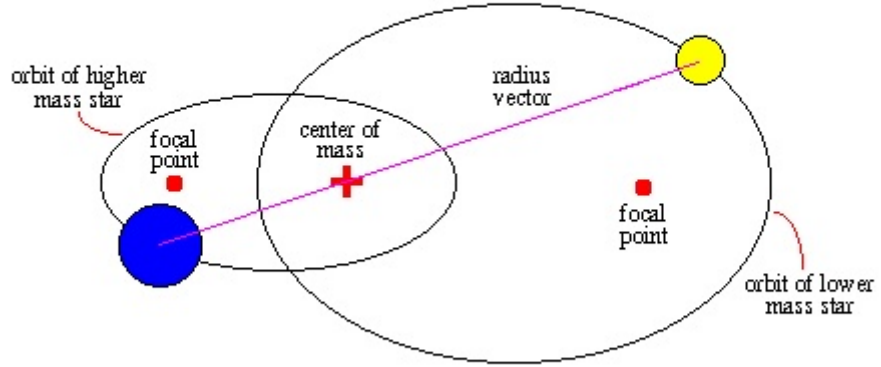


Figure 10.1 Illustration of the properties of a visual binary system.

A key difference from the case of a satellite orbiting the earth, or a planet orbiting a star, is that in binary stars, the masses can become comparable. In this case, each star (1,2) now moves around the *center of mass* at a fixed distance a_1 and a_2 , with their ratio given by $a_2/a_1 = M_1/M_2$ and their sum by $a_1 + a_2 = a$. In terms of the full separation, the orbital distance of, say, star 1 is thus given by

$$a_1 = a \frac{M_2}{M_1 + M_2} . \quad (10.2)$$

For the given period P , the associated orbital speeds for star 1 is given by $V_1 = 2\pi a_1/P$. For a stable, circular orbit, the outward centrifugal force on star 1,

$$F_{c1} = \frac{M_1 V_1^2}{a_1} = \frac{4\pi^2 M_1 a_1}{P^2} = \frac{4\pi^2 a}{P^2} \frac{M_1 M_2}{M_1 + M_2} , \quad (10.3)$$

must balance the gravitational force from eqn. (10.1), yielding

$$\frac{GM_1 M_2}{a^2} = \frac{4\pi^2 a}{P^2} \frac{M_1 M_2}{M_1 + M_2} . \quad (10.4)$$

This can be used to obtain the sum of the masses,

$$M_1 + M_2 = \frac{4\pi^2}{G} \frac{a^3}{P^2} = \frac{a_{au}^3}{P_{yr}^2} M_{\odot} , \quad (10.5)$$

where the latter equality shows that evaluating the distance in au and the period

in years gives the mass in units of the solar mass. For a visual binary in which we can actually see both stars, we can separately measure the two orbital distances, yielding then the mass ratio $M_2/M_1 = a_1/a_2$. The mass for, e.g., star 1 is thus given by

$$M_1 = \frac{a_{au}^3}{(1 + a_1/a_2) P_{yr}^2} M_\odot. \quad (10.6)$$

The mass for star 2 can likewise be obtained if we just swap subscripts 1 and 2.

Equations (10.5) and (10.6) are actually forms of Kepler's 3rd law for planetary motion around the Sun. Setting $M_1 = M_\odot$ and the planetary mass M_2 , we first note that for all planets the mass is much smaller than for the Sun, $M_2/M_1 = a_1/a_2 \ll 1$, implying that the Sun only slightly wobbles (mostly to the counter the pull of the most massive planet, namely Jupiter), with the planets thus pretty much all orbiting around the Sun. If we thus ignore M_2 and plug in $M_1 = M_\odot$ in eqn. (10.5), we recover Kepler's third law in (almost) the form in which he expressed it,

$$P_{yr}^2 = a_{au}^3. \quad (10.7)$$

To be precise, Kepler showed that in general the orbits of the planets are actually *ellipses*, but this same law applies in that case if we replace the circular orbital distance a with the “semi-major axis” of the ellipse. A circle is just a special case of an ellipse, with the semi-major axis just equal to the radius.

In general, of course, real binary systems often have elliptical orbits, which, moreover, lie in planes that are not always normal to the observer line of sight. These systems can still be fully analyzed using the elliptical orbit form of Newton's generalization of Kepler's 3rd law, as derived, e.g. in Ch. 4 of the Astro 45 notes by Bill Press:

<http://www.lanl.gov/DLDSTP/ay45/ay45c4.pdf>

Indeed, by watching the rate of movement of the stars along the projected orbit, the inclination effect can even be disentangled from the ellipticity.

10.2 Spectroscopic binaries

As noted, there are many stellar binary systems in which the angular separation between the components is too close to readily resolve visually. However, if the orbital plane is not perpendicular to the line of sight, then the orbital velocities of the stars will give a variable Doppler shift to each star's spectral lines. The effect is greatest when the orbits are relatively *close*, and in a plane *containing* the line of sight, conditions which make such *spectroscopic binaries* complement the wide visual binaries discussed above. Figure 10.2 illustrates the basic features of a spectroscopic binary system.

If the two stars are not too different in luminosity, then observations of the combined stellar spectrum show spectral line signatures of both stellar spectra.

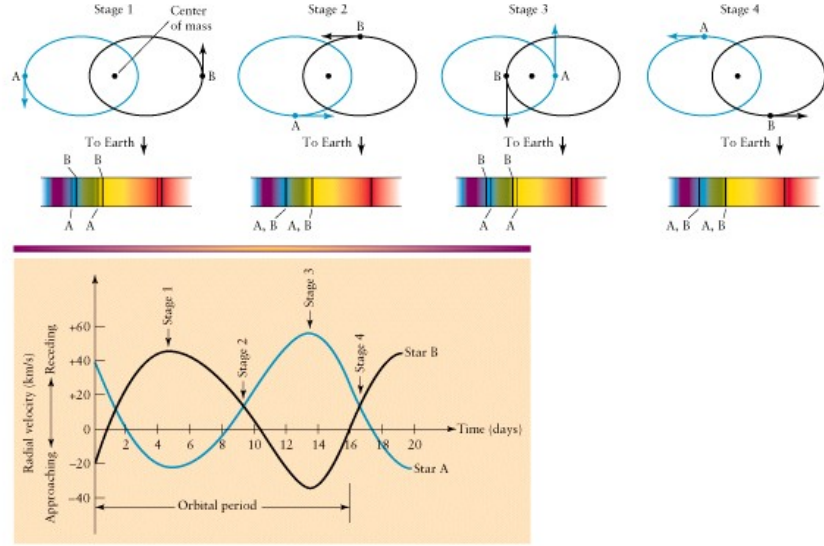


Figure 10.2 Illustration of the periodic Doppler shift of spectral lines in a spectroscopic binary system.

As the stars move around each other in such “double-line” spectroscopic binaries¹, the changing Doppler shift of each of the two spectral line patterns provides information on the changing orbital velocities of the two components, V_1 and V_2 .

Considering again the simple case of circular orbits but now in a plane *containing* the line of sight, the inferred radial velocities vary sinusoidally with semi-amplitudes given by the orbital speeds $V_1 = 2\pi a_1/P$ and $V_2 = 2\pi a_2/P$, where a_1 and a_2 are the orbital radii defined earlier. Since the period P is the same for both stars, the ratio of these inferred velocity amplitudes gives the stellar mass ratio, $M_1/M_2 = V_2/V_1$. Using the same analysis as used for visual binaries, but noting now that $a = a_1 + a_2 = PV_1(1 + V_2/V_1)/2\pi$, we obtain a “velocity form” of Kepler’s third law given in eqn. (10.6)²,

$$M_1 = \frac{1}{2\pi G} V_2^3 P (1 + V_1/V_2)^2$$

$$M_1 = \left[\frac{V_2}{V_e} \right]^3 P_{yr} (1 + V_1/V_2)^2 M_\odot, \quad (10.8)$$

¹ In “single-line” spectroscopic binaries, the brighter “primary” star is so much more luminous that the lines of its companion are not directly detectable; but this secondary star’s presence can nonetheless be inferred from the periodic Doppler shifting of the primary star’s lines due to its orbital motion.

² In the case that the orbital axis is inclined to the line of sight by an angle i , then these scalings generalize with a factor $\sin^3 i$ multiplying the mass, with the velocities representing the inferred Doppler shifted values.

where the latter equality gives the mass in solar units when the period is evaluated in years, and the orbital velocity in units of the Earth's orbital velocity, $V_e = 2\pi \text{ au/yr} \approx 30 \text{ km/s}$. Again, an analogous relation holds for the other mass, M_2 , if we swap indices 1 and 2.

10.3 Eclipsing binaries

In some (relatively rare) cases of close binaries, the two stars actually pass in front of each other, forming an eclipse that temporarily reduces the amount of light we see. Such eclipsing binaries are often also spectroscopic binaries, and the fact that they eclipse tells us that the inclination of the orbital plane to our line of sight must be quite small, implying that the Doppler shift seen in the spectral lines is indeed a direct measure of the stellar orbital speeds, without the need to correct for any projection effect. Moreover, observation of the eclipse intervals provides information that can be used to infer the individual stellar radii.

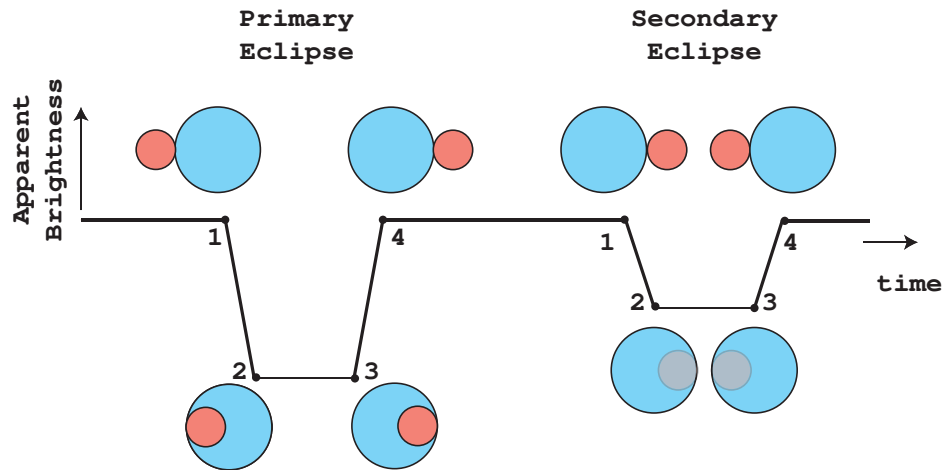


Figure 10.3 Illustration of the how the various contact moments of eclipsing binary star system correspond to features in the observed light curve.

Consider, for example, the simple case that the orbital plane of the two stars is *exactly* in our line of sight, so that the centers of the two pass directly over each other. As noted the maximum Doppler shifts of the lines for each star then gives us a direct measure of their respective orbital speeds, V_1 and V_2 . In our above simple example of circular orbits, this speed is constant over the orbit, including during the time when the two stars are moving across our line of sight, as they pass into and out of eclipse. In eclipse jargon, the times when the stellar rims just touch are called “contacts”, labeled 1-4 for first, second, etc. Clearly then, once the stellar orbital speeds are known from the Doppler shift, then the radius

of the smaller star (R_2) can be determined from the time difference between the first (or last) two contacts,

$$R_2 = (t_2 - t_1)(V_1 + V_2)/2. \quad (10.9)$$

Likewise, the larger radius (R_1) comes from the time between the second and fourth (or third to first) contacts,

$$R_1 = (t_4 - t_2)(V_1 + V_2)/2. \quad (10.10)$$

In principal, one can also use the other, weaker eclipse for similar measurements of the stellar radii.

Of course, in general, the orbits are elliptical and/or tilted somewhat to our line of sight, so that the eclipses don't generally cross the stellar centers, but typically move through an off-center chord, sometimes even just grazing the stellar limb. In these cases information on the radii requires more complete modeling of the eclipse, and fitting the observations with a theoretical light curve that assumes various parameters. Indeed, to get good results, one often has to relax even the assumption that the stars are spheres with uniform brightness, taking into account the mutual tidal distortion of the stars, and how this affects the brightness distributions across their surfaces. Such details are somewhat beyond the scope of this general survey course (but could make the basis for an interesting term paper or project).

10.4 Mass-Luminosity scaling from astrometric and eclipsing binaries

In the above simple introduction of the various types of binaries, we've assumed that the orientation, or "inclination" angle i , of the binary orbit relative to our line of sight is optimal for the type of binary being considered, i.e. looking face on – with $i = 0$ inclination between our sight line and the orbital axis – for the case visual binaries; or edge-on – with $i = 90^\circ$ – for spectroscopic binaries in which we wish to observe the maximum Doppler shift from the orbital velocities. Of course, in practice binaries are generally at some intermediate, often unknown inclination, leaving an ambiguity in the determination of the mass (typically scaling with $\sin^3 i$) for a given system.

Fortunately, in the relatively few binary systems that are both spectroscopic (with either single or double lines) and either astrometric or eclipsing, it becomes possible to determine the inclination, and so unambiguously infer the *masses* of the stellar components, as well as the *distance* to the system. Together with the observed apparent magnitudes, this thus also gives the associated luminosities of these stellar components.

Figure 10.4 plots $\log L$ vs. $\log M$ (in solar units) for a sample of such astrometric (blue) and eclipsing (red) binaries, showing a clear trend of increasing luminosity with increasing mass. Indeed, a key result is that for many of the stars

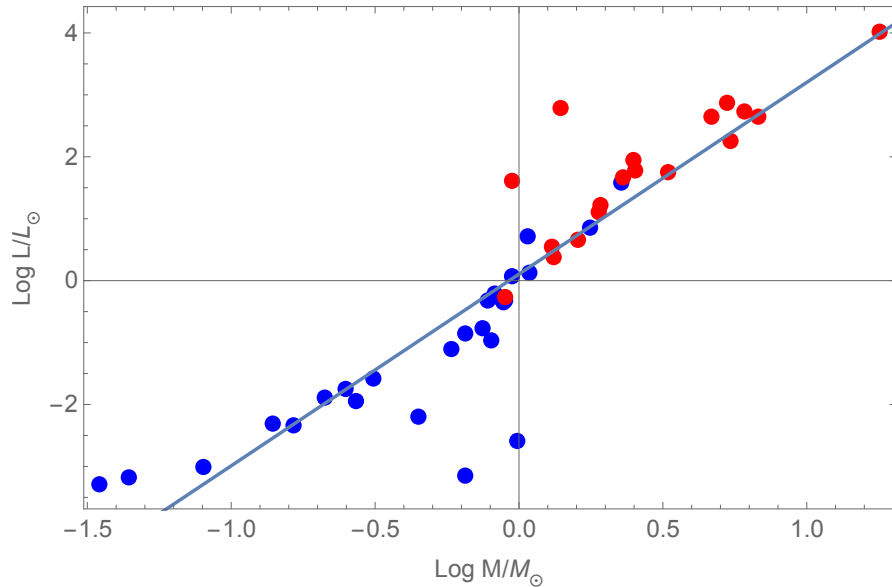


Figure 10.4 A log-log plot of luminosity vs. mass (in solar units) for a sample of 26 astrometric (blue, lower points) binaries and 18 double-line eclipsing (red, upper points) binaries. The best-fit line shown follows the empirical scaling, $\log(L/L_\odot) \approx 0.1 + 3.1 \log(M/M_\odot)$.

(typically those on the main sequence), the data can be well fit by a straight line in this log-log plot, implying a power-law relation between luminosity and mass,

$$\boxed{\frac{L}{L_\odot} \approx \left(\frac{M}{M_\odot}\right)^{3.1}}. \quad (10.11)$$

Part II of these notes will use the stellar structure equations for hydrostatic equilibrium and radiative diffusion to explain why the luminosities of main sequence stars roughly follow this observed scaling with the cube of the stellar mass, $L \sim M^3$ (see §17.1).

10.5 Questions and Exercises

Quick Question 1: Note that the net amount of stellar surface eclipsed is the same whether the smaller or bigger star is in front. So why then is one of the eclipses deeper than the other? What quantity determines which of the eclipses will be deeper?

QQ 2:

Over a period of 10 years, two stars separated by an angle of 1 arcsec are observed

to move through a full circle about a point midway between them on the sky. Suppose that over a single year, that midway point is observed itself to wobble by 0.2 arcsec due to the parallax from Earth's own orbit.

- a. How many pc is this star system from earth?
- b. What is the physical distance between the stars, in au.
- c. In solar masses, what are the masses of each star, M_1 and M_2 .

11 Inferring Stellar Rotation

Let us conclude our discussion of stellar properties by considering ways to infer the rotation of stars. All stars rotate, but in cool, low-mass stars like the Sun the rotation is quite slow, with for example the Sun having a rotation period $P_{rot} \approx 26$ days, corresponding to an equatorial rotation speed $V_{rot} = 2\pi R_{\odot}/P_{rot} \approx 2$ km/s. In hotter, more-massive stars, the rotation can be more rapid, typically 100 km/s or more, with some cases (e.g., the Be stars) near the “critical” rotation speed at which material near the equatorial surface would be in a Keplerian orbit! While the rotational evolution of stars is a topic of considerable research interest, its importance is generally of secondary importance compared to, say, the stellar mass.

11.1 Rotational broadening of stellar spectral lines

In addition to the Doppler shift associated with the star’s overall motion toward or away from us, there can be a *differential* Doppler shift from the parts of the star moving toward and away as the star rotates. This leads to a *rotational broadening* of the spectral lines, with the half-width given by

$$\frac{\Delta\lambda_{rot}}{\lambda_o} \equiv \frac{V_{rot} \sin i}{c}, \quad (11.1)$$

where V_{rot} is the stellar surface rotation speed at the equator, and $\sin i$ corrects for the inclination angle i of the rotation axis to our line of sight. If the star happens to be rotating about an axis pointed toward our line of sight ($i = 0$), then we see no rotational broadening of the lines. Clearly, the greatest broadening is when our line of sight is perpendicular to the star’s rotation axis ($i = 90^\circ$), implying $\sin i = 1$, and thus that $V_{rot} = c\Delta\lambda_{rot}/\lambda_o$.

Figure 11.1 illustrates this rotational broadening. The left-side schematic shows how a rotational broadened line profile for flux vs. wavelength takes on a *hemi-spherical*¹ form. For a rigidly rotating star, the line-of-sight component of

¹ If flux is normalized by the continuum flux F_c , then making the plotted profile actually trace a hemi-sphere requires the wavelength to be scaled by $\lambda_n \equiv \Delta\lambda_{rot}/r_o$, where $\Delta\lambda_{rot}$ and r_o are the line’s rotational half-width and central depth, defined respectively by eqns. (11.1) and (11.3).

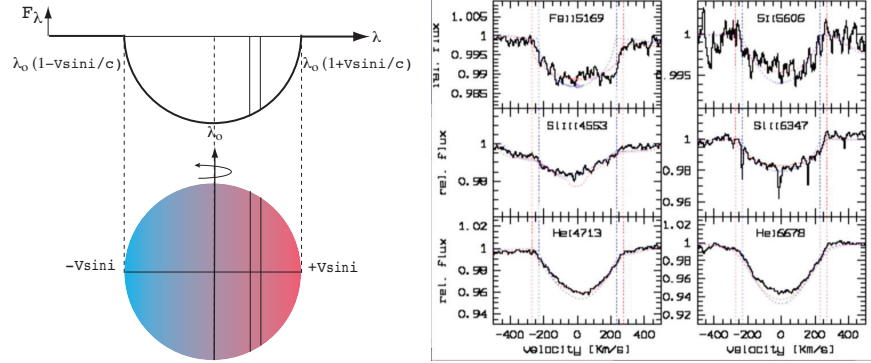


Figure 11.1 Left: Schematic showing how the Doppler shift from rigid body rotation of a star (bottom) – with constant line-of-sight velocity along strips parallel to the rotation axis – results in a hemi-spherical line-absorption-profile (top). Right: Observed rotational broadening of lines for a sample of stars with (quite rapid) projected rotation speeds $V \sin i > 100$ km/s.

the surface rotational velocity just scales in proportion to the apparent displacement from the projected stellar rotation axis. Thus for an intrinsically narrow absorption line, the total amount of reduction in the observed flux at a given wavelength is just proportional to the *area* of the vertical strip with a line-of-sight velocity that Doppler-shifts line-absorption to that wavelength. As noted above, the total width of the profile is just twice the star’s projected equatorial rotation speed, $V \sin i$.

The right panel shows a collection of observed rotationally broadened absorption lines for a sample of quite rapidly rotating stars, i.e. with $V \sin i$ more than 100 km/s, much larger than the ~ 1.8 km/s rotation speed of the solar equator. The flux ratio here is relative to the nearby “continuum” outside the line.

Note that the reduction at line-center is typically only a few percent. This is because such rotational broadening preserves the total amount of reduced flux, meaning then that the relative depth of the reduction is diluted when a rapid apparent rotation significantly broadens the line.

A convenient measure for the total line absorption is the “equivalent width”,

$$W_\lambda \equiv \int_0^\infty \left(1 - \frac{F_\lambda}{F_c}\right) d\lambda, \quad (11.2)$$

which represents the width of a “saturated rectangle” with same integrated area of reduced flux. For a line with equivalent width W_λ and a rotationally broadened half-width $\Delta\lambda_{rot}$, the central reduction in flux is just

$$r_o \equiv 1 - \frac{F_{\lambda_o}}{F_c} = \frac{2}{\pi} \frac{W_\lambda}{\Delta\lambda_{rot}}. \quad (11.3)$$

For example, for the He 471.3 nm line plotted in the left, lowermost box in the right panel of figure 11.1, the central reduction is just $r_o \approx 1 - 0.96 =$

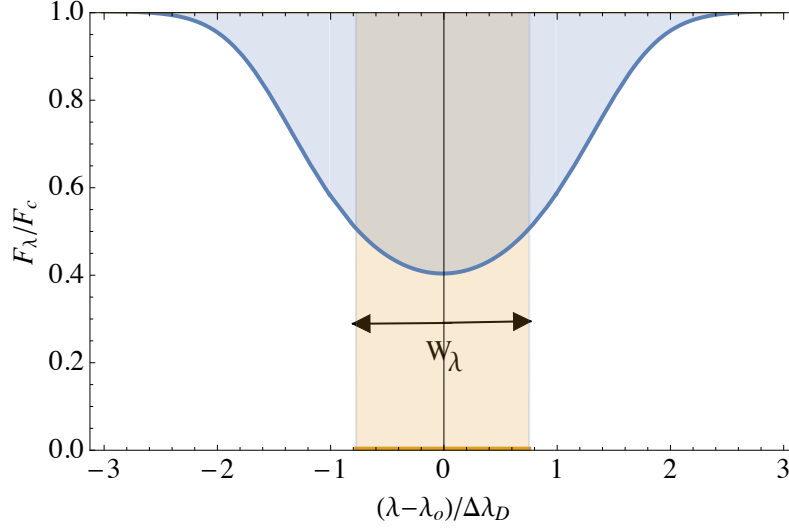


Figure 11.2 Illustration of the definition of the wavelength equivalent width W_λ . The blue curve plots the wavelength variation of the residual flux (relative to the continuum, i.e., F_λ/F_c) for a sample absorption line, with the shaded blue area illustrating the total fractional reduction of continuum light. The tanned plot show a box profile with width W_λ , defined such that the total tanned area is the same the blue area for the line profile.

f

0.04, while the velocity half-width (given e.g. by the vertical red-dotted lines) is $V \sin i \approx 275$ km/s, corresponding to a wavelength half-width $\Delta\lambda_{rot} \approx 0.43$ nm. This implies an equivalent width $W_\lambda \approx 0.027$ nm, or about 17 km/s in velocity units.

11.2 Rotational period from starspot modulation of brightness

When Galileo first used a telescope to magnify the apparent disk of the Sun, he found it was not the “perfect orb” idealized from antiquity, but instead had groups of relatively dark “sunspots” spread around the disk. By watching the night-to-night migration of these spots from the east to west, he could see directly that the Sun is rotating, with a mean period² of about 25 d.

Though other stars are too far away to directly resolve the stellar disk and thus make similar direct detections of analogous “starspots”, in some cases such spots are large and isolated enough that careful photometric measurement of the apparent stellar brightness shows a regular modulation over the stellar rotation period P .

² Actually, the Sun does not rotate as a rigid-body, but has about 10% faster rotation at its equator than at higher latitudes.

11.3 Questions and Exercises

Quick Question 1: A line with rest wavelength $\lambda_o = 500 \text{ nm}$ is rotational broadened to a full width of 0.5 nm . Compute the value of $V \sin i$, in km/s .

Exercise 1: Derive eqn. (11.3) from the definitions of rotational Doppler width $\Delta\lambda_{rot}$ (11.1) and equivalent width W_λ (11.2), using the wavelength scaling given in footnote 1.

If the star also shows rotationally broadened spectral lines with an associated inferred projected rotational speed $V_{rot} \sin i$, then the basic relation $V_{rot} = 2\pi R/P$ implies a constraint on the minimum possible value for the stellar radius, $R_{min} = V_{rot} \sin i P / 2\pi$.

12 Light Intensity and Absorption

12.1 Intensity vs. Flux

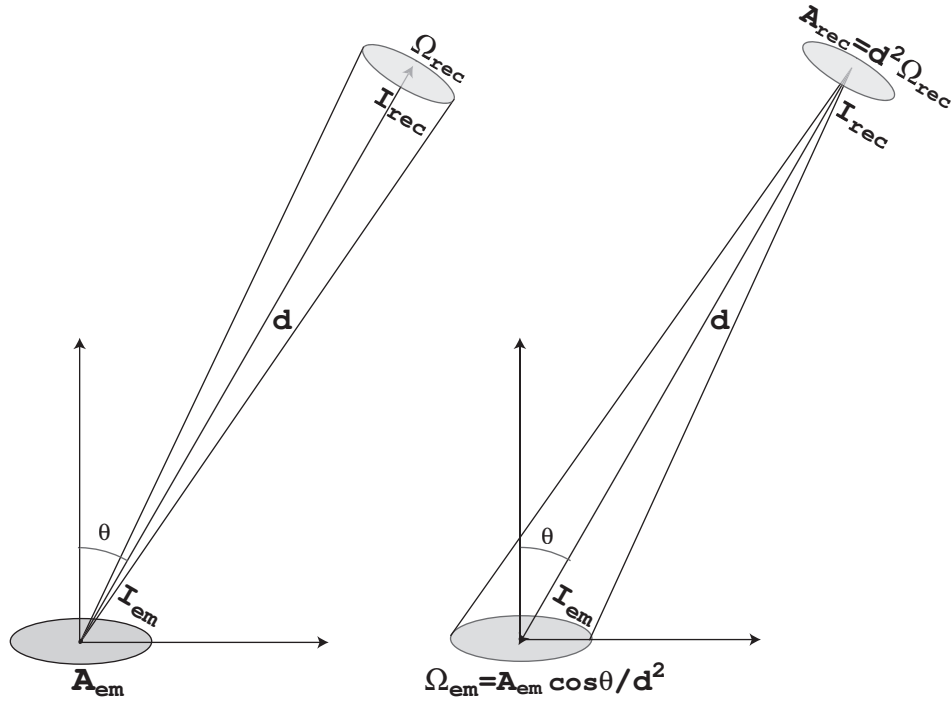


Figure 12.1 Left: The intensity I_{em} emitted into a solid angle Ω_{rec} located along a direction that makes an angle θ with the normal of the emission area A_{em} . Right: The intensity I_{rec} received into an area $A_{rec} = d^2 \Omega_{rec}$ at a distance d from the source with projected solid angle $\Omega_{em} = A_{em} \cos \theta / d^2$. Since the emitted and received energies are equal, we see that $I_{em} = I_{rec}$, showing that intensity is invariant with distance d .

Our initial introduction of surface brightness characterized it as a flux confined within an observed solid angle, F/Ω . But actually the surface brightness is directly related to a more general and fundamental quantity known as the

*Specific*¹ *Intensity* I . In the exterior of stars, the intensity is set by the surface brightness $I \approx F/\Omega$, but it can also be specified in the stellar *interior*, where it characterizes the properties of the radiation field as energy generated in the core is transported to the surface.

A simple analog on Earth would be an airplane flying through a cloud. Viewed from outside, the cloud has a surface brightness from reflected sunlight, but as the plane flies into the cloud, the light becomes a “fog” coming from all directions, with the specific intensity in any given direction depending on the details of the scattering through the cloud.

Formally, intensity is *defined* as the radiative energy per unit area and time that is pointed into a specific patch of solid angle $d\Omega$ centered on a specified *direction*. The left side of figure 12.1 illustrates the basic geometry. As the solid angle of the projected *emitting* area declines with the inverse square of the distance, $\Omega_{em} = A_{em} \cos \theta / d^2$, the fixed solid angle *receiving* the intensity *grows* in area in proportion to the distance-squared, $A_{rec} = \Omega_{rec} d^2$. In essence, the two distances cancel, and so the *intensity remains constant with distance*.

In this context it is perhaps useful to think of intensity in terms of a narrow *beam* of light in a particular direction – like a laser “beam” –, whereas the flux depends on just the total amount of light energy that falls on a given area of a detector, regardless of the original direction of all the individual “beams” that this might be made up of. However, while valid, this perspective might suggest that intensity is a vector and flux a scalar, whereas in fact the *opposite* is true. The intensity has a directional *dependence* through the specification of the direction of the solid angle being emitted into, but it itself is a scalar! The flux measures the rate of energy (a *scalar*) through a given area, but this has an associated direction given by the *normal* to that surface area; thus the flux is a *vector*, with its three components given by the three possible orientations of the normal to the detection area.

For stars in which the emitted radiation is, at least to a first approximation, spherically symmetric, the only non-zero component of the flux is along the radial direction away from the star. If the angle between any given intensity beam I with the radial direction is written as θ , then its contribution to the radial flux is proportional to $I \cos \theta$; the total radial flux is then obtained by integrating this contribution over solid angle,

$$F = \int I(\theta) \cos \theta d\Omega = 2\pi \int_0^\pi I(\theta) \cos \theta \sin \theta d\theta. \quad (12.1)$$

The latter equality applies the spherical coordinate form for solid angle, integrated over the azimuthal coordinate (ϕ) to give the factor 2π .

As a simple example, let us assume the Sun has a surface brightness I_\odot that is constant, both over its spherical surface of radius R_\odot , and also for all *outward* directions from the surface.² Now consider the flux $F(d)$ at some distance d (for

¹ Often this “Specific” qualifier is dropped, leaving just “Intensity”.

² Actually, the light from the Sun is “limb darkening”, meaning the intensity directly

example at Earth, for which $d = 1$ au). At this distance, the visible solar disk has been reduced to a half-angle $\theta_d = \arcsin(R_\odot/d)$, so that the angle range for the non-zero local intensity has shrunk to the range $0 < \theta < \theta_d$, i.e.

$$\begin{aligned} I(\theta) &= I_\odot \quad ; \quad 0 < \theta < \theta_d \\ &= 0 \quad ; \quad \theta_d < \theta < \pi \end{aligned} \quad (12.2)$$

Noting that $\cos \theta_d = \sqrt{1 - R_\odot^2/d^2}$, we then see that evaluation of the integral in eqn. (12.1) gives for the flux

$$F(d) = \pi I_\odot (1 - \cos^2 \theta_d) = \pi I_\odot \frac{R_\odot^2}{d^2}. \quad (12.3)$$

Again, within the cone of half-angle θ_d around the direction toward the Sun's center, the observed intensity is the same as at the solar surface $I = I_\odot$. But the shrinking of this cone angle with distance gives the flux an inverse-square dependence with distance, $F(d) \sim 1/d^2$.^s To obtain the flux at the surface radius R of a blackbody, we note that $I = B(T)$ for *outward* directions with $0 < \theta < \pi$, but is zero for inward directions with $\pi/2 < \theta < \pi$. Noting then that $\sin \theta d\theta = -d \cos \theta$, we can readily carry out the integral in eqn. (12.1), yielding then the Stefan-Boltzmann law (cf. eqn. 5.2) for the radially outward surface flux

$$F_* \equiv F(R) = \pi B(T) = \sigma_{sb} T^4. \quad (12.4)$$

This also follows from the general flux scaling given in eqn. (12.3) if we just set $d = R_\odot$ and $I_\odot = B(T)$.

12.2 Absorption mean-free-path and optical depth

The light we see from a star is the result of competition between thermal emission and absorption by material within the star. Let us first focus on the basic scalings for the absorption by considering the simple case of a beam of intensity I_o along a direction z perpendicular to a planar layer that consists of a local number density $n(z)$ of absorbing particles of projected *cross sectional area* σ (see figure 12.2.) We can characterize the *mean-free-path* that light can travel before being absorbed within the layer as

$$\ell \equiv \frac{1}{n \sigma} = \frac{1}{\rho \kappa}. \quad (12.5)$$

The latter equality instead uses the *mass* density $\rho = \mu n$, where μ is the mean mass of stellar material per absorbing particle. The cross section divided by this mass defines what's called the *opacity*, $\kappa \equiv \sigma/\mu$, which is thus simply the cross section per unit mass of the absorbing medium.

^supward is greater than that at more oblique angles toward to the local horizon, or limb.
See Appendix D.2.

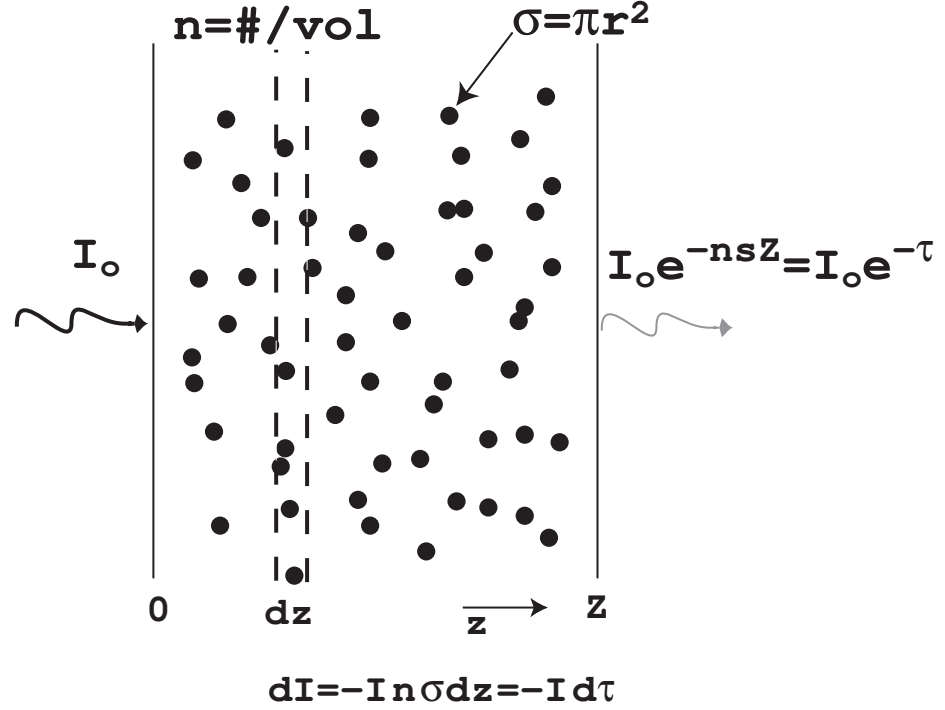


Figure 12.2 Illustration of the attenuation of an intensity beam I_o by a planar layer of absorbing particles with cross section σ and number density n .

Within a narrow (differential) layer between z and $z + dz$, the probability of light being absorbed is just $d\tau \equiv dz/\ell$. This implies an associated *fractional* reduction $dI/I = -d\tau$ in the local intensity $I(z)$. We can thus write this change in intensity in terms of a simple differential equation,

$$\frac{dI}{dz} = -\kappa \rho I \quad \text{or} \quad \frac{dI}{d\tau} = -I. \quad (12.6)$$

Straightforward integration using the boundary condition $I(z = 0) = I_o$ at the layer's leading edge at $z = 0$ gives

$$\boxed{I(z) = I_o e^{-\tau(z)}}, \quad (12.7)$$

where

$$\tau(z) \equiv \int_0^z \frac{dz'}{\ell} = \int_0^z n(z') \sigma dz' = \int_0^z \kappa \rho(z') dz' \quad (12.8)$$

represents the integrated *optical depth* from the surface to a position z within the layer. It is clear from the initial definition that one can think of optical depth as simply the number of mean-free-paths between two locations.

12.3 Inter-stellar extinction and reddening

One practical example of such exponential reduction of light by absorption is the case of inter-stellar “extinction” of starlight. The space between stars – called the *Inter-Stellar Medium* (ISM) – is not completely empty, but contains a certain amount of gas and dust. Compared to a stellar atmosphere, or indeed even to a strong terrestrial vacuum, the density is very small, often only a few atoms per cubic centimeter, or a few hundred dust particles per cubic *kilometer*. But over the huge distances between stars, the associated optical depth τ for extinction of the star’s light by scattering and/or absorption can become quite significant, leading to a substantial reduction in the star’s apparent brightness.

For a star of radius R and surface intensity I , the luminosity is $L = 4\pi^2 R^2 I$, and in the absence of any absorption the *intrinsic* flux at a distance d is just $F_{\text{int}}(d) = L/4\pi d^2 = \pi I(R/d)^2$. But in the case with ISM absorption, the *observed* flux is again (cf. eqn. 12.7) reduced by the optical depth exponential absorption factor

$$F_{\text{obs}}(d) = F_{\text{int}}(d)e^{-\tau}, \quad (12.9)$$

where the subscripts stand for “*observed*” and “*intrinsic*”. The level of this ISM absorption can also be characterized in terms of the number of *magnitudes of extinction*,

$$A \equiv m_{\text{obs}} - m_{\text{int}} = 2.5 \log \left(\frac{F_{\text{int}}}{F_{\text{obs}}} \right) = 2.5 \tau \log e \approx 1.08 \tau. \quad (12.10)$$

In interpreting the observed magnitude of a “standard candle” star with known luminosity, the failure to account for any such extinction can lead to an inferred distance d_{inf} that *overestimates* the star’s true distance d . For observations in the visual band V, we can define an associated visual extinction $A_V \equiv V_{\text{obs}} - V_{\text{int}} \approx 1.08\tau_V$, where τ_V is the optical depth within the visual band.

In practice, interstellar extinction is generally dominated by the opacity associated with interstellar grains of dust. For large dust grains, the absorption cross section just depends on the physical size, for example given by $\sigma = \pi r^2$ for spherical grains of radius r .

But interstellar dust grains are often very tiny, even microscopic, with sizes of less than a micron, and so comparable to the wavelength of optical light. For light in the red or infra-red that has a wavelength larger than the dust size, $\lambda > r$, the effective cross section, and thus the associated dust opacity, is *reduced*, because, in a loose sense, the dust particle can only interact with a fraction of the light wave. Because this redder, longer wavelength light is less strongly absorbed than the bluer, shorter wavelengths, the remaining light tends to appear “*reddened*”, much in the same way as the Sun’s light at sunset.

This reddening can be quantified in terms of a formal *color excess*, defined in

terms of the standard B and V filters of the Johnson photometric system,

$$E_{B-V} \equiv (B - V)_{\text{obs}} - (B - V)_{\text{int}}. \quad (12.11)$$

This color excess tends to increase with increasing visual extinction magnitude $A_V \equiv V_{\text{obs}} - V_{\text{int}}$. If the intrinsic colors are known (e.g., from the star's spectral type), then, for a given model of the wavelength dependence of the opacity, measuring this color excess makes it possible to estimate of the visual extinction magnitude A_V . Among other things, this allows one to reduce or remove the error in determining the stellar distance.

The detailed variation of dust opacity depends on the size, shape, and composition of the dust, but often it is approximated as scaling as an inverse power law in wavelength, i.e.

$$\kappa(\lambda) \sim \lambda^{-\beta},$$

where the power index (a.k.a. “reddening exponent”) ranges from $\beta \approx 1$ for “Mie scattering” to $\beta \approx 4$ for “Rayleigh scattering”.

The latter is a good approximation for scattering by air molecules and dust in the earth's atmosphere. The scattering of blue light out of the direction from the Sun makes the sunset red, while all that scattered blue light makes the sky blue.

For ISM dust, the weaker $\beta \approx 1$ scaling is more appropriate, but even this can make a marked difference in the level of extinction for different wavelengths. Details are discussed in §21.3 on dust extinction by Giant Molecular Clouds of the ISM.

12.4 Questions and Exercises

Quick Question 1

- (a.) Suppose spherical dust grains have a radius $r = 0.1 \text{ cm}$ and individual mass density $\rho_g = 1 \text{ g/cm}^3$. What is their cross section σ , mass m , and associated opacity κ ?
- (b.) If the number density of these grains is $n_d = 1 \text{ cm}^{-3}$, what is the mass density of dust ρ_d and the mean free path ℓ for light?
- (c.) What is the optical depth at a physical depth 1 m into a planar layer of such dust absorbers?
- (d.) What fraction of impinging intensity I_o makes it to this depth?

Quick Question 2: Derive expressions for d_{inf}/d in terms of both the absorption magnitude A and the optical depth τ .

13 Observational Methods

13.1 Telescopes as light buckets

Stars are so far away that, of the several hundred *billion* in our galaxy, only about 5000 are visible to our naked eye. Even when the pupils in our eye are dark adapted, they have a maximum diameter of only about 7 mm, limiting the light reaching our retina. Telescopes provide a way to greatly improve on this by collecting the light from a much greater aperture, effectively acting as “light buckets”. For a circular aperture of diameter D , the amount of light gathered scales in proportion to the collection area,

$$A = \frac{\pi}{4} D^2. \quad (13.1)$$

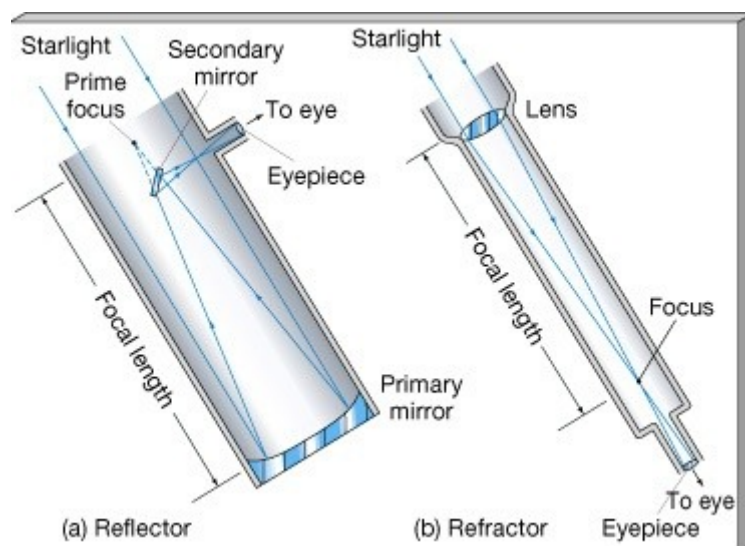


Figure 13.1 Illustration of basic differences between a refracting vs. reflecting telescopes.

As illustrated in figure 13.1, telescopes generally can be categorized as *refractors* vs. *reflectors*. Much like our eyes, refractor telescopes use a lens to bend

incoming light into a focus; but such lenses can have a diameter up to about 100 times larger than our pupils, thus collecting 10,000 times as much light. For even larger lenses, housing the long focal length becomes unwieldy, and so this is near the practical upper limit for refractor telescopes.

But reflector telescopes, wherein light is collected by large primary mirror, can be built with much larger total apertures. The largest optical reflectors currently in operation have diameters about 10 meters¹, formed by combining ~ 100 meter-size hexagonal mirror segments. For example, a 7 m diameter mirror that is now a thousand times the aperture of our pupil can collect a *million* times as much light as our eyes.

Using the definition of magnitude from §3.3, this can be used to derive a general formula for the increase in limiting magnitude resulting just from an increased aperture,

$$m_{\text{lim}} \approx 7.5 + 5 \log(D/\text{cm}). \quad (13.2)$$

Eqn. (13.2) does apply directly to amateur telescopes that are used to view the sky by eye. But in practice, large research telescopes use modern digital camera detectors with efficiencies that well exceed that of our retina. By also integrating the exposure for many minutes or hours, they can detect much fainter objects with magnitudes much larger than the aperture-based limit (13.2). In practice, the limit is often set by the background darkness of the local sky, one reason modern telescopes are built at remote sites, well away from the light pollution of cities.

Quick Question 1:

The human eye has an integration time $t_{\text{int}} \approx 0.1$ sec, and a photon detection efficiency $\epsilon \approx 0.1$. Generalize equation (13.2) to estimate the m_{lim} for a telescope detector with higher values of t_{int} and ϵ .

13.2 Angular resolution

Another advantage to a large mirror diameter is that it enables a higher angular resolution. For light of wavelength λ , the diffraction from a telescope with diameter D sets a fundamental limit to the smallest possible angular separation that can be resolved,

$$\alpha = 1.22 \frac{\lambda}{D} = 0.25 \text{ arcsec} \frac{\lambda/\mu\text{m}}{D/\text{m}} = 2.5 \text{ arcsec} \frac{\lambda/\text{cm}}{D/\text{km}}, \quad (13.3)$$

where the latter two equalities are scaled respectively for optical and radio telescopes.

For ground-based optical telescopes, this ideal diffraction limit is not generally reached, because turbulence in Earth's atmosphere blurs the image over

¹ The Extremely Large Telescope (ELT) currently under construction in Atacama desert in Chile will have diameter of 39 meters! First light is planned for 2025.

~ 1 arcsec or more, an effect known as “astronomical seeing”. But this can be reduced to resolutions approaching 0.1 arcsec through a technique called *adaptive optics*, wherein reflection from a laser beam shot up into the sky is used to estimate these seeing distortions, and then dynamically deform secondary mirrors to correct for them.

The sharpest focus requires the primary mirror to have a *parabolic* shape. The primary mirror of the Hubble Space Telescope (HST) was mistakenly (and quite infamously) ground instead to a spherical form, leading then to a “spherical aberration” in images that had to be subsequently corrected by secondary optics. But with this correction, and despite the modest 2.4 m diameter of its primary mirrors, HST’s location in orbit above atmospheric distortions and light pollution has helped it revolutionize observational astronomy².

Since radio waves can propagate even through the clouds that block visible light, large radio telescopes have been constructed even in locations with poor weather conditions. The radio reflector is now called a ‘dish’, with the largest ones (e.g., the 300-meter Arecibo telescope in Puerto Rico) built into natural depressions in the terrain, extending over hundreds of meters. Such dishes are not steerable, but by positioning the receiver around the focal plane they can effectively aim at a range of positions within 30° from the local zenith. The largest steerable dishes range up to 100 m in diameter.

The Very Large Array (VLA) in New Mexico consists of 27 individual dishes that are each 25 m in diameter, positioned on tracks that can spread them over a baseline of up to 30 km. While the sensitivity is set by the combined collective area of the many dishes, a technique called *interferometry* combines their signals to give angular resolution associated with this wider baseline. An extension of this technique, called Very Long Baseline Interferometry (VLBI) can even combine signals from telescopes spread all around the globe; their diffraction limit can thus in principle approach that of a telescope the size of the entire Earth!

An impressive recent example is the Event Horizon Telescope (EHT), which used an array of two dozen telescopes to image the mm-wavelength emission around a black hole, with angular resolution near 25 *micro*-arcsec! The Atacama Large Millimeter Array (ALMA) consists of 66 antennas spread over up to 16 km of the very dry Atacama desert in Chile; the limited water vapor reduces the absorption of mm and sub-mm waves enough to allow detection in this intermediate waveband, which is key for, e.g., diagnosing conditions in star-forming regions that have many magnitudes of extinction (sections 12.3, 21.3, 22) at shorter wavelengths in the visible.

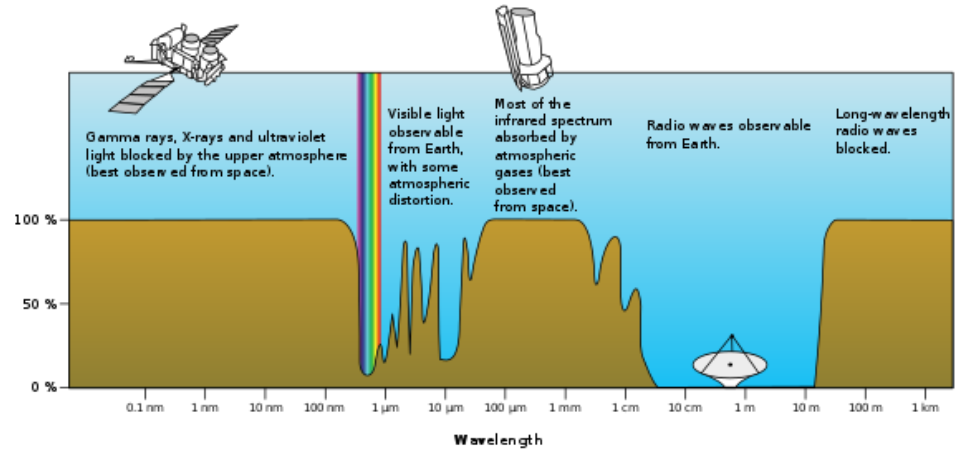


Figure 13.2 The percentage of electromagnetic radiation that is blocked by Earth's atmosphere, plotted as function of wavelength. Image credit: NASA.

13.3 Space-based missions

More generally, as illustrated in figure 13.3, Earth's atmosphere effectively *blocks* radiation in some spectral bands, e.g., at shorter wavelengths ($\lambda < 350$ nm) below the visible. Thus observations in these UV, X-ray, and gamma-ray regions can only be done from orbiting space-based platforms above the atmosphere.

The Hubble telescope has been a principal instrument in the near UV ($100 \text{ nm} < \lambda < 400 \text{ nm}$), allowing improved study of hot stars and warm interstellar gas with temperatures $10,000 \text{ K} < T < 100,000 \text{ K}$. (In the far and extreme UV ($10 \text{ nm} < \lambda < 92.1 \text{ nm}$), ionization by Hydrogen in the local interstellar medium largely attenuates radiation from any more distant sources).

At X-ray wavelengths ($0.10 \text{ nm} < \lambda < 10 \text{ nm}$), corresponding to high-energy photons ($0.1 \text{ keV} < E < 100 \text{ keV}$), telescopes probe very energetic regions with temperatures heated to millions of Kelvin, e.g. from accretion onto compact objects like neutron stars and black holes (§20.5), or hot interstellar bubbles that are shock heated by supernova explosions (§21.2).

At still shorter, gamma-ray wavelengths $\lambda < 0.01 \text{ nm}$, with still higher photon energies ($E > \text{MeV}$), detectors have discovered mysterious gamma ray bursts. The longer duration ($> \text{few sec}$) ones are now thought to arise from “hypernovae” associated with collapse of rotating cores of massive stars, while the shorter-duration bursts are understood to originate from the “kilonovae” associated with merger of neutron stars (§20.6).

Finally, orbiting telescopes have also been used for the part of the infrared

² Particularly noteworthy are the weeklong exposures allowed by its uniquely dark sky background; known as the Hubble Deep Fields, these exposures revealed huge numbers of very faint, very distant galaxies up to 10 Gly away.

blocked by the atmosphere. Such infrared observation are particularly key to studying cool dense regions of the interstellar medium where dust absorption leads to many magnitudes of extinction in visible light; these are often regions of active star formation (sections 12.3, 21.3, 22).

A full list of space-based telescopes is given at:

https://en.wikipedia.org/wiki/List_of_space_telescopes.

13.4 Questions and Exercises

Quick Question 1: a. The VLA works at radio frequencies 1-30 GHz. Work out associated wavelength range in cm. b. Then for $D=30$ km, work out the associated angular resolutions α , in arcsec.

14 Our Sun

Thus far our discussion of stellar properties has mainly used our Sun as a benchmark for key overall quantities, like surface temperature, radius, mass, and luminosity. But of course the close proximity of the Sun, and its extreme apparent brightness, makes it by far the most important star for our lives here on Earth. Other stars are so far away that even to our most powerful telescopes they appear as mere points of light, from which we can only measure the overall flux, or apparent brightness. But the Sun is close enough that we can resolve its *surface brightness*, or intensity, across its angular diameter of about 0.5° . When its extreme brightness is suitably filtered by a dark lens, it appears to our eyes as a generally featureless disk. But even with his small, primitive telescope, Galileo was able to discover darkened blemishes we now call sunspots, and so disprove the classical ideal of the Sun as a perfect, heavenly sphere.

In modern times we have access to powerful telescopes, both on the ground and in space, that observe and monitor the Sun over a wide range of wavelength bands. These vividly demonstrate that the Sun is in fact highly structured and variable over a wide range of spatial and temporal scales, and so provide a sobering reality check on our own simple idealizations of stars as being constant, featureless, spherically symmetric balls of gas.

14.1 Imaging the solar disk

Figure 14 shows images of the solar disk made by NASA's orbiting Solar Dynamics Observatory (SDO) in 13 different wavebands, chosen to highlight different layers of the solar atmosphere, corresponding to the labeled temperatures.

In the top row the third image from the left shows the standard visual continuum, often dubbed 'white-light', formed in the layer known as the *photosphere*. As detailed in Appendix D.2, the less-bright, redder intensity toward the disk edge, known as *limb darkening*, results from the vertical decline of temperature through this photospheric layer. The sunspots below and to the left of the disk center appear dark in this visual image because, as shown in the 'magnetogram' just to its left¹, these are regions of strong magnetic field, with the light to dark

¹ Such magnetograms detect the circular and linear polarization of light induced by magnetic fields on the solar surface.

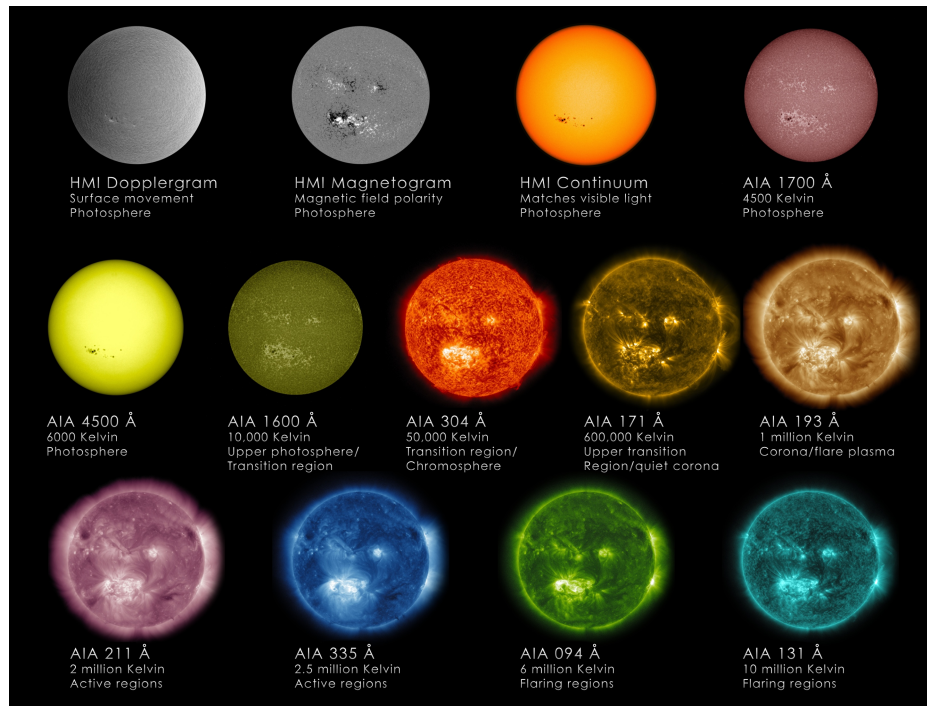


Figure 14.1 Full-disk images of the Sun at 13 different wavelengths, made by the NASA’s Solar Dynamics Observatory (SDO). This includes images from both from the Advanced Imaging Assembly (AIA), which helps show how solar material moves around the Sun’s atmosphere, and the Helioseismic and Magnetic Imager (HMI), which focuses on the movement and magnetic properties of the Sun’s surface. Each wavelength was chosen to highlight a particular part of the Sun’s atmosphere, from the solar photosphere, through the chromosphere, and up to the upper reaches of the corona. Credits: NASA/SDO/Goddard Space Flight Center. Further details at: <https://www.nasa.gov/content/goddard/how-sdo-sees-the-sun>

switch indicating a change in the magnetic polarity; the fields are so strong that they inhibit the convective transport of energy from below, thus making sunspots relatively cool, and thus darker.

But these fields also are conduits for magnetic waves and turbulence, which when dissipated at higher layers actually add extra mechanical heating that cause the temperature in these upper layers to *rise*! Instead of the effective temperature $T \approx 5800\text{ K}$ that characterizes the photosphere, there first develops a hotter *chromosphere*, with temperature in the range $10,000 - 50,000\text{ K}$. This is followed by an abrupt jump across a narrow *transition region* to temperatures of *millions* of Kelvin (!) in the solar *corona*.

In the more-opaque UV wavebands that are formed in these higher layers, the regions above sunspots, known as *active regions*, are thus actually *brighter* than

the surrounding areas. For example, in the central panel of the middle row, which is tuned to 304 \AA emission from ionized Helium at temperatures of 50,000 K in the upper chromosphere, the active regions are bright, though there is still emission over the entire solar disk. But as one moves to the *far* UV and X-ray diagnostics (right middle and bottom row) that are formed at the MK temperatures of the corona, the contrast becomes greater, with some nearly dark regions that have little or no emission, known as *coronal holes*.

Figure 14.1 provides a schematic summary of these various layers and features of the solar atmosphere.

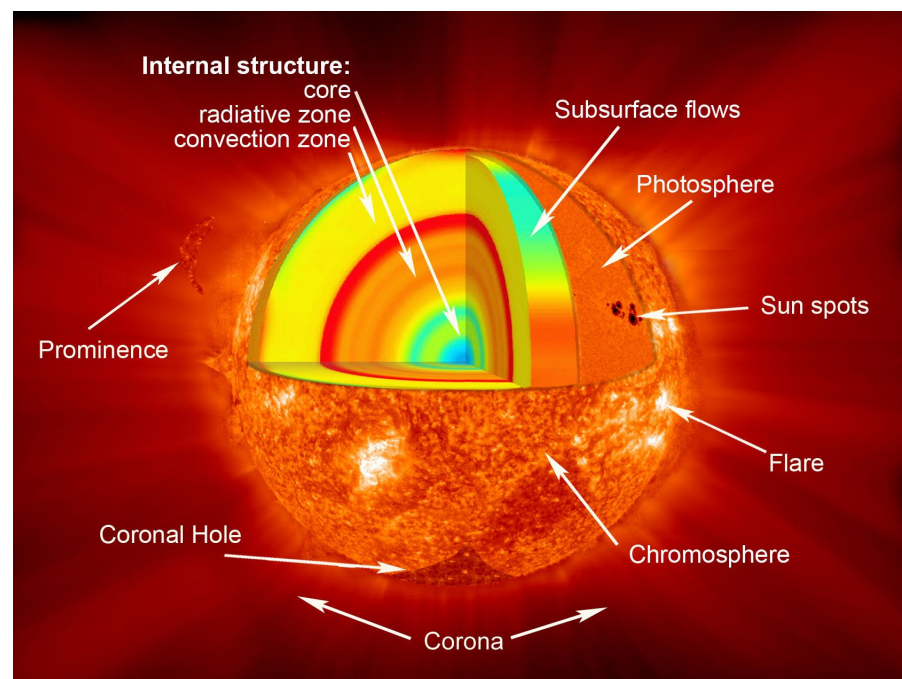


Figure 14.2 Schematic summary of key regions and features of the solar atmosphere, along with interior cutout showing the Sun's nuclear-burning core, intermediate radiative diffusion region, and near-surface convection zone. Image credit: NASA

14.2 Corona and solar wind

Though very hot, the corona has a very low density, even above active regions. At visual wavelengths it is thus nearly transparent, and so generally hard to see. Fortunately, by an amazing coincidence, Earth's moon has nearly the same angular size as the Sun, and so in rare and brief instances, there occurs a *solar eclipse*, during which the moon just covers up the bright solar disk. As shown in

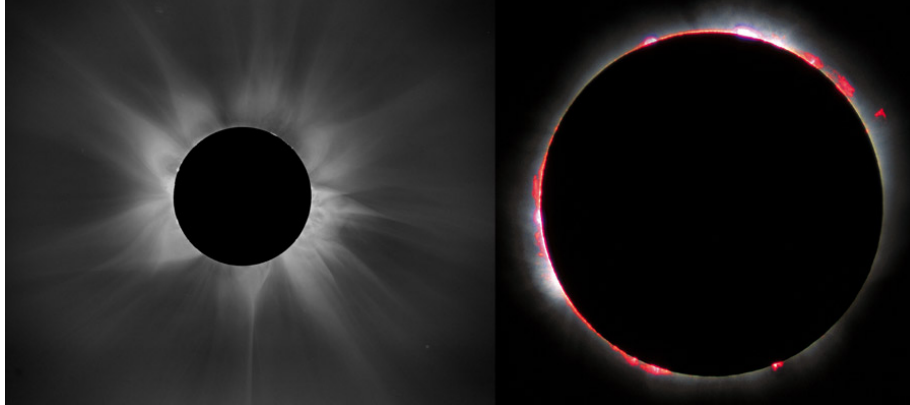


Figure 14.3 Left: Eclipse image of the solar corona during time of extensive active regions on the sun. Right: Image of the red Hydrogen emission from chromospheric active regions during the same eclipse. Credit: NCAR’s High Altitude Observatory (corona) and Luc Viatour (chromosphere)

the left panel of figure 14.1, this allows us to see the corona as visible solar light scattered by electrons in the corona’s tenuous, but highly ionized gas. In the right panel, the rim of red light comes from the chromosphere², via the magnetic suspension of hot gas in active regions, leading to Hydrogen Balmer- α ($n = 3$ to $n = 2$; see Appendix A.) emission at the red wavelength 6563 \AA . The magnetic fields from these active regions rise up into the corona, forming closed magnetic loops that connect footpoints of opposite magnetic polarity on the surface.

The corona is so hot that the Sun’s gravity cannot, by itself, keep the gas bound against a pressure-driven outward expansion known as the *solar wind*. But in regions with closed magnetic loops, the magnetic field tension holds the gas back against this expansion, allowing such regions to keep a high pressure and density, and thus making them more visible in both white light and X-ray signatures.

Coronal holes arise in *open* field regions between such closed loops, allowing the gas to escape into the outward solar wind expansion. This gives then a lower coronal density and relatively low brightness in both scattered white-light, and X-ray emission. The coronal magnetic field is thus the key cause of the coronal structure seen in figure 14.1.

The radial streamers³ at the tops of the coronal loops show that wind expansion wins out in the outer corona, effectively pulling open the closed field lines there. The resulting solar wind expands outward, past the Earth and even all the other planets, extending to distances $> 100 \text{ au}$, until it is finally stopped

² This red color led to the name ‘chromosphere’, from the Greek *chroma* for color.

³ Sometimes referred to as “helmet streamers”, due their resemblance to German WWI army helmets.

by running into the local interstellar medium. The full region within this wind-termination boundary is referred to as the *heliosphere*.

As illustrated below in figure 24.2, the magnetosphere formed by Earth's own magnetic field shields our planet and its atmosphere from a direct hit by the solar wind, instead just channeling any solar wind plasma toward the magnetic poles, where interaction with the atmosphere forms the aurora, a.k.a. the norther and souther lights. In contrast, the lack of a strong field on Mars has allowed the solar wind to gradually erode its now much thinner atmosphere. As discussed in section 24.3, this can affect the habitability of extra-solar planets around cool stars with coronal winds.

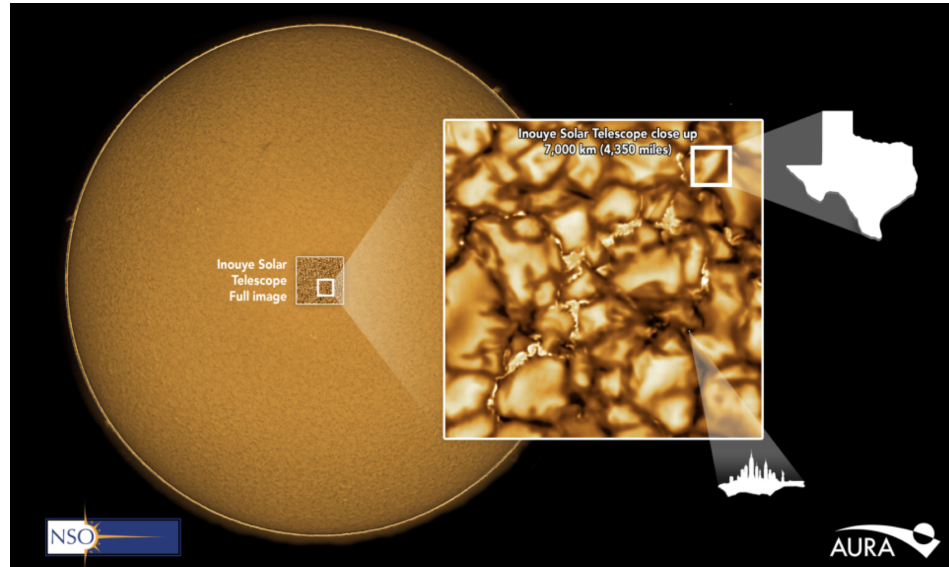


Figure 14.4 Illustration of how the DKIST telescope allows us to zoom in to image structure on the solar surface down to a resolution of ~ 0.1 arcsec, or ~ 70 km, only about twice the size of Manhattan. The irregular granulation structures have typical size of ~ 2 arcsec, or ~ 1500 km, about the size of Texas. They represent cells of convection, with brighter center upwelling from the hotter interior, bounded by narrow lanes of cooler darker downflows. Image credit: NSO/NSF/AURA; visit www.nso.edu

14.3 Convection as a driver of solar structure and activity

The angular diameter of the solar disk is

$$\alpha_{\odot} = \frac{2R_{\odot}}{\text{au}} \approx 0.01 \text{ rad} \approx 0.5^{\circ} \approx 1800 \text{ arcsec}. \quad (14.1)$$

This means the roughly 1 arcsec resolution limit from atmospheric seeing allows for about 1800 resolution elements across the solar disk, representing a physical size of $s = \text{au} \times \text{arcsec} \approx 2R_{\odot}/1800 \approx 700 \text{ km}$. With special techniques to correct for atmospheric seeing, it is possible to reach a factor 10 higher resolution, so down to 0.1 arcsec, or a physical size $s \approx 70 \text{ km}$.

As illustrated in figure 14.2, such resolution is achieved by DKIST⁴, the currently most advanced ground-based solar telescope. Its primary mirror has a diameter $D = 4 \text{ m}$, which from eqn. (13.3) gives a diffraction-limit resolution $< 0.1 \text{ arcsec}$ in the visible. Its site at an altitude of about 3000m atop the Haleakala volcano on the island of Maui, Hawaii was chosen for its relatively stable air and so good, sub-arcsec seeing, which with adaptive-optics correction allows resolution that approaches this diffraction limit.

Zooming in to a small segment of the disk, figure 14.2 shows the Sun's *granulation* pattern, with central bright cells bounded by narrow, darker lanes. This is characteristic of a systematic gas motion called *convection*. Hotter gas in the interior wells upward in the cell centers, making them hotter and thus brighter. After this gas cools by radiation into space, it falls back downward in the narrow lanes that bound the cells, which being cooler also appear darker. As detailed below in section 17.3, such convection arises in the near-surface layers of relatively cool stars like the Sun, from the blocking of radiative diffusion by the enhanced opacity associated with ionization of neutral Hydrogen. An animation of the dynamical variation of this convective structure is given in <https://www.youtube.com/watch?v=4nieF-e000s>.

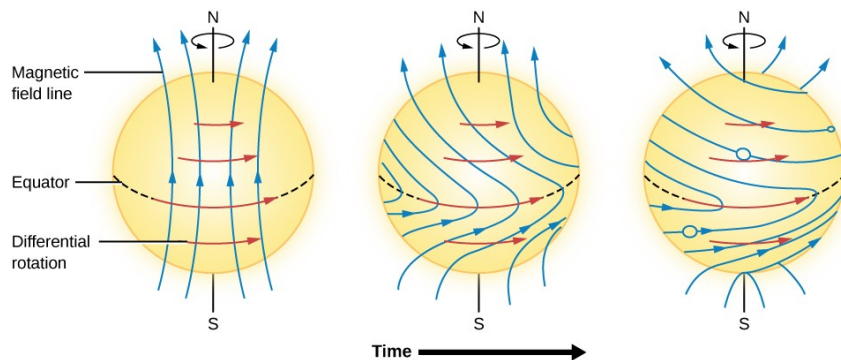


Figure 14.5

Such convection combines with the Sun's rotation to generate magnetic fields through a "rotation-convection magnetic dynamo". Although hydrogen gas in the solar atmosphere is mostly neutral, other elements lose a sufficient number of their less tightly bound electrons to make the overall gas behave like an

⁴ For *Daniel K. Inouye Solar Telescope*, named in honor of the Hawaii senator who championed funding for the project in the US congress.

ionized plasma, with a high electrical conductivity. Such a conducting plasma makes any magnetic field “frozen-in”, or effectively stuck to the local plasma. Near and below the stellar surface, where the gas energy density dominates over that associated with the magnetic field, the stretching and compression of any embedded magnetic field acts to amplify that field. For example, in granulation convection cells, the dense material upwelling in the center tends to sweep any magnetic field lines to the cell edges, thus concentrating the field in the narrow dark lanes. The small-scale bright regions in the dark lanes in figure 14.2 are sites of such locally concentrated magnetic field.

Above the solar surface, the rapid decrease in gas density and pressure with height means that in the upper layers of the Sun’s atmosphere, in the *chromosphere* and extending up into surrounding *corona*, it is the magnetic field that dominates and channels the gas, leading to the extensive coronal structure shown in the eclipse image in figure 14.1.

Finally, as illustrated in figure 14.3, larger-scale interior generation of magnetic fields occurs through the interaction of convection with the Sun’s *differential rotation*. The latter refers to the fact that the Sun does not rotate as a solid body, like the Earth or any planet, but instead actually has a faster angular rotation (shorter rotation period) at its equator than at higher latitudes towards its poles. As field lines are stretched azimuthally, they eventually form kinks that pop up through the solar photosphere, forming sunspot pairs with opposite magnetic polarity. With increased strength and complexity of the field, coupled with foot-point wandering induced by convection, can lead to localized regions of *magnetic reconnection*. The sudden release of magnetic energy leads to a localized *flare* in brightness in wavebands from the visible to X-ray, as shown in several panels on the bottom row of figure 14. Over time this reconnection dissipation leads to an overall decline in magnetic field strength and complexity, and so an associated decline in solar activity till this reaches a relatively quiescent minimum, whereupon the cycle restarts with winding up of the large-scale residual field by differential rotation. This is the origin of the 11-year cycle seen in, e.g. sunspot number, as well as other signatures of solar activity.

Through monitoring spectroscopic signatures of activity, including coronal X-ray emission, activity cycles have been inferred in other cool, solar type stars, albeit with varying periods ranging from about a year to many decades. This illustrates again how the Sun provides us with benchmark for complex structure and activity in stars that we only see as points of light, reminding that they too are far more complex than our idealized steady, spherically models would imply.

14.4 Questions and Exercises

Quick Question 1:

Part II

Stellar Structure & Evolution

15 Hydrostatic Balance between Pressure and Gravity

We have seen in part I how a star's color or peak wavelength λ_{max} indicates its characteristic temperature near the stellar surface. But what about the temperature in the star's deep interior? Intuitively, we expect this to be much higher than at the surface, but under what conditions does it become hot enough to allow for nuclear fusion to power the star's luminosity? And how does it scale quantitatively with the overall stellar properties, like mass M , radius R , and perhaps luminosity L ?

To answer these questions, let us identify two distinct considerations for our intuition that the interior temperature should be much higher than at the surface.

The first we might characterize as the “blanketing” by the overlying layers, which traps any energy generated in the interior, much as a blanket in bed traps our body heat, keeping our skin temperature at a comfortable warmth, instead of the relative chilliness of having it exposed to open air. In this picture, the equilibrium interior temperature depends on the rate of energy generation (from metabolism for our bodies, or nuclear fusion for stars) and the ‘insulation thickness’ of the overlying of material to the surface (given by the optical depth; see §16.)

But distinct from this consideration of the *transport of energy* from the interior, there is for a star a dynamical requirement for *force* or *momentum* balance, to keep the star supported against the inward pull of its own self-gravity. Since stars are gaseous, without the tensile strength of a solid body, this gravitational support is supplied by increased internal gas *pressure* P , allowing the star to remain in a static equilibrium. This high gas pressure arises from a combination of high density and high temperature. As detailed in §15.3, this allows us to determine a characteristic interior temperature, through a further application of the Virial theorem for bound systems that was briefly discussed for bound orbits in section 7.4 of part I.

15.1 Hydrostatic equilibrium

To quantify this gravitational equilibrium for a static star, consider, as illustrated in figure 15.1, a thin radial segment of thickness dr with local density ρ and downward gravitational acceleration g . The mass-per-unit-area of this layer

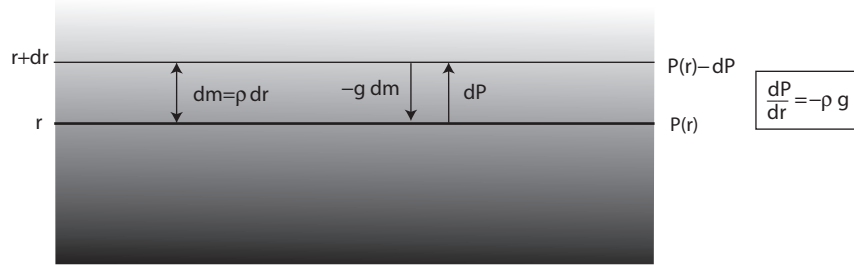


Figure 15.1 Illustration of the radial decline of gas pressure P due to local mass density ρ and downward gravity g .

is $dm = \rho dr$, with corresponding weight-per-unit-area $g dm$. To support this weight, the gas pressure at the lower end of this layer must be higher by amount $|dP| = g dm = \rho g dr$ than the upper end, implying

$$\boxed{\frac{dP}{dr} = -\rho g}, \quad (15.1)$$

a condition known as *Hydrostatic Equilibrium*.

For an *ideal gas*, the pressure depends on the product of the number density $n = \rho/\bar{\mu}$ and temperature T ,

$$\boxed{P = nkT = \rho \frac{kT}{\bar{\mu}} \equiv \rho c_s^2}, \quad (15.2)$$

where $k = 1.38 \times 10^{-16}$ erg/K is Boltzmann's constant, $\bar{\mu}$ is the average mass – a.k.a. the “mean molecular weight” – of all particles (i.e., both ions and electrons) in the gas, and the final equation defines the isothermal¹ sound speed, $c_s \equiv \sqrt{kT/\bar{\mu}}$.

For any given element, the *fully ionized* molecular weight is just set by the nuclear mass Am_p (because the electron mass is by comparison negligible) divided by the nuclear charge number plus one (for the one nucleus + Z_n electrons that balance the nuclear charge eZ_n), $\mu = m_p A/(Z_n + 1)$. For a gas mixture with mass fraction X , Y , and Z for H, He, and metals, the overall mean molecular weight is then obtained by a weighted average of the *inverses* ($m_p/\mu = (Z_n + 1)/A$) of the individual components (as in a parallel circuit), yielding

$$\bar{\mu} = \frac{m_p}{2X + 3Y/4 + Z/2} \approx 0.6m_p \equiv \bar{\mu}_\odot, \quad (15.3)$$

¹ This speed, which was first derived by Newton, would only be the speed of sound if the gas remained strictly constant temperature (isothermal). In practice, the temperature fluctuations associated with the gas compressions make the actual “adiabatic” speed of sound slightly higher, by a factor $\sqrt{\gamma}$, where γ is the ratio of specific heats (5/3 for a monatomic gas).

where the last equality is for the solar case with $X = 0.72$, $Y = 0.26$, and $Z = 0.02$. More generally, for fully ionized gases the proton-mass-scaled molecular weight $\bar{\mu}/m_p$ can range from $1/2$ for pure H ($X = 1$), to $4/3$ for pure He ($Y = 1$), to a maximum of 2 for pure heavy metals ($Z = 1$).

15.2 Pressure scale height and thinness of surface layer

The ratio of eqns. (15.2) to (15.1) defines a characteristic *pressure scale height*,

$$H \equiv \frac{P}{|dP/dr|} = \frac{kT}{\bar{\mu}g} = \frac{c_s^2}{g}, \quad (15.4)$$

where the absolute value of the pressure gradient dP/dr (which itself is negative) ensures the scale height is positive.

At the stellar surface radius $r = R$, where the gravity and temperature approach their fixed surface values $g_* = GM/R^2$ and $T = T_*$, the scale height becomes quite small, typically only a tiny fraction of the stellar radius,

$$\frac{H}{R} = \frac{kT_*/\bar{\mu}}{GM/R} = \frac{2c_{s*}^2}{V_{esc}^2} \approx 0.0005 \frac{T_*/T_\odot}{\bar{\mu}/\bar{\mu}_\odot} \frac{R/R_\odot}{M/M_\odot}, \quad (15.5)$$

where V_{esc} is the escape speed introduced in part I. For the solar atmosphere, the sound speed is $c_{s*} \approx 9$ km/s, about $1/60$ th of the surface escape speed $V_{esc} = 620$ km/s.

If we further idealize a stellar atmosphere as being roughly *isothermal*, i.e. with a nearly constant temperature $T \approx T_*$, then, since the gravity is also effectively fixed at the surface value, we see that the scale height also becomes constant. This makes it easy to integrate the hydrostatic equilibrium equation (15.1), thus giving the variation of density and pressure in terms of a simple exponential stratification with height $z \equiv r - R$,

$$\frac{P(z)}{P_*} = \frac{\rho(z)}{\rho_*} = e^{-z/H}, \quad (15.6)$$

where the asterisk subscripts denote values at some surface layer where $z \equiv 0$ (or $r = R$). In practice the temperature variations in an atmosphere are gradual enough that quite generally both pressure and density very nearly follow such an exponential stratification.

The results in this section actually apply to *any* gravitationally bound atmosphere, not only for stars but also for planets, including the earth, with similarly small characteristic values for the ratio H/R . This is the basic reason that the earth's atmosphere is confined to such a narrow layer around its solid surface, meaning that at just a couple hundred kilometers altitude it is nearly a vacuum, so tenuous that it imparts only a weak drag on orbiting satellites.

For stars or gaseous giant planets without a solid surface, it means that the dense, opaque regions have only a similarly narrow transition to the fully transparent upper layers, thus giving them a similarly sharp visual edge as a solid

body. For stars it means that models of the escape of interior radiation through this narrow atmospheric layer can essentially ignore the stellar radius, allowing the emergent spectrum to be well described by a planar atmospheric model fixed by just two parameters – surface temperature and gravity – and not dependent on the actual stellar radius.

15.3 Hydrostatic balance in stellar interior and the virial temperature

This hydrostatic balance must also apply in the stellar interior, but now both the temperature and gravity have a strong spatial variation. At any given interior radius r , the local gravitational acceleration depends only on the mass *within* that radius,

$$M(r) \equiv 4\pi \int_0^r \rho(r') r'^2 dr'. \quad (15.7)$$

This thus requires the hydrostatic equilibrium equation to be written in the somewhat more general form,

$$\boxed{\frac{dP}{dr} = -\rho(r) \frac{GM(r)}{r^2}}. \quad (15.8)$$

This represents one of the key equations for stellar structure.

The implications of hydrostatic equilibrium for the hot interior of stars are quite different from the steep exponential pressure drop near the surface; indeed they allow us now to derive a remarkably simple scaling relation for a characteristic interior temperature T_{int} .

For this consider the associated interior pressure P_{int} at the *center* of the star ($r = 0$); to drop from this high central pressure to the near-zero pressure at the surface, the pressure gradient averaged over the whole star must be $|dP/dr| \sim P_{int}/R$. We can similarly characterize the gravitational attraction in terms of the surface gravity $g_* = GM/R^2$ times an interior density that scales as $\rho_{int} \sim P_{int}\bar{\mu}/kT_{int}$. Applying these in the basic definition of scale height (15.4), we find that for the interior $H \approx R$, which in turn implies for this characteristic stellar interior temperature,

$$\boxed{T_{int} \approx \frac{GM\bar{\mu}}{kR} \approx 14 \times 10^6 \text{K} \frac{M/M_\odot}{R/R_\odot}}. \quad (15.9)$$

Thus, while surface temperatures of stars are typically a few thousand Kelvin, we see that their interior temperatures are typically of order 10 *million* Kelvin! As discussed below (§18), this is indeed near the temperature needed for nuclear fusion of Hydrogen into Helium in the stellar core.

This close connection between thermal energy of the interior ($\sim kT$) to the star's gravitational binding energy ($\sim GM\bar{\mu}/R$) is really just another example of the *Virial theorem* for gravitationally bound systems, as discussed in part I for

the case of bound orbits. The temperature is effectively a measure of the average kinetic energy associated with the random *thermal* motion of the particles in the gas. Thermal energy is thus just a specific form of kinetic energy, and the Virial theorem tells us that the average kinetic energy in a bound system equals one-half the magnitude of the gravitational binding energy.

15.4 Questions and Exercises

Quick Question 1:

- (a.) For a typical temperature on a spring day ($\sim 50^\circ\text{F}$), compute the scale height H for the earth (in km), and its ratio to earth's radius, H/R_e .
- (b.) Relative to values at sea level, compute the pressure and density at a typical height $h = 300$ km for an orbiting satellite.

Quick Question 2:

- (a.) Compute the escape speed (in km/s) from stars with $M = M_\odot$ and $R = 2R_\odot$; with $M = 2M_\odot$ and $R = R_\odot$.
- (b.) For these stars, estimate the associated central temperature.

Quick Question 3:

- (a.) For a constant density stellar envelope, show that the mass within radius r is given by $M(r) = M(R)(r/R)^3$, where R is the stellar radius.
- (b.) For such a star, show by explicit integration of the hydrostatic equilibrium equation (15.8) that the core temperature is $T(0) = 7$ MK, and so half the value given by eqn. (15.9).

16 Transport of Radiation from Interior to Surface

16.1 Random walk of photon diffusion from stellar core to surface

Let us next turn to the “blanketing” effect of the star’s material in trapping the heat and radiation of the interior. Within a star the absorption of light by stellar material is *counteracted by thermal emission*. As illustrated in figure 16.1, radiation generated in the deep interior of a star undergoes a diffusion between multiple encounters with the stellar material before it can escape freely into space from the stellar surface. The number of mean-free-paths ℓ from the center at $r = 0$ to the surface at radius $r = R$ now defines the central optical depth

$$\tau_c = \int_0^R \frac{dr'}{\ell} = \int_0^R \kappa \rho dr'. \quad (16.1)$$

As discussed in Appendix C, the opacity in stellar interiors typically has a CGS value of $\kappa \approx 1 \text{ cm}^2/\text{g}$, with a minimum set by the value for Thomson scattering by free electrons, $\kappa_e \approx 0.34 \text{ cm}^2/\text{g}$. We can then estimate a typical value of this central optical depth by simply taking the density to be roughly characterized by its volume average $\bar{\rho} = M/(4\pi R^3/3)$, where again M and R are the stellar mass and radius. For the Sun this works out to give $\bar{\rho}_\odot \approx 1.4 \text{ g/cm}^3$, i.e. just above the density of water. (The Sun wouldn’t quite float in your bathtub.) Since the opacity is also near unity in CGS units, the average mean-free-path for scattering in the Sun is just $\bar{\ell}_\odot \approx 0.7 \text{ cm}$.

In the core of the actual Sun, the density is typically a hundred times higher than this mean value, so the core mean-free-path is a factor hundred smaller, i.e. $\ell_{\text{core}} \approx 0.07 \text{ mm}$! But either way, the mean-free-path is much, much smaller than the solar radius $R_\odot \approx 700,000 \text{ km} = 7 \times 10^{10} \text{ cm}$. This implies the optical depth from the center to surface is truly enormous, with a typical value

$$\tau_c \approx \frac{R_\odot}{\bar{\ell}_\odot} \approx 10^{11}. \quad (16.2)$$

The total number of scatterings needed to diffuse from the center to the surface can then be estimated from a basic “*random walk*” argument. The simple 1D version states that after N left/right random steps of unit length, the root-mean-square (rms) distance from the origin is \sqrt{N} . For the 3D case of stellar diffusion, this rms number of unit steps can be roughly associated with the total number of mean-free-paths between the core and surface, i.e. $\sqrt{N} \approx \tau$. This implies that

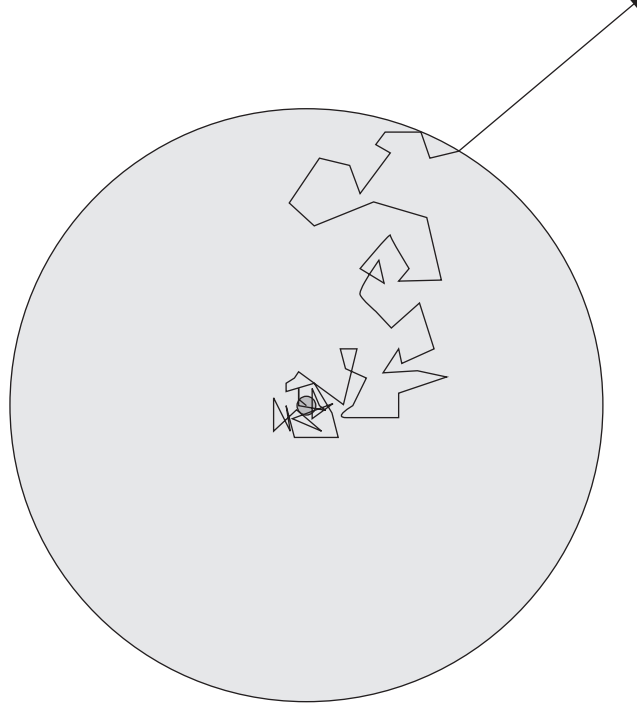


Figure 16.1 Illustration of the random-walk diffusion of photons from the core to surface of a star.

photons created in the core of the Sun need to scatter a total of $N \approx \tau^2 \approx 10^{22}$ times to reach the surface!

In traveling from the Sun's center to its surface, the *net* distance is just the Sun's radius R_\odot ; but the *cumulative* path length traveled is *much* longer, $\ell_{tot} \approx N\bar{\ell}_\odot \approx \tau^2\bar{\ell}_\odot \approx \tau R_\odot$. For photons traveling at the speed of light $c = 3 \times 10^{10}$ cm/s, the total time for photons to diffuse from the center to the surface is thus

$$t_{diff} = \tau^2 \frac{\bar{\ell}_\odot}{c} \approx \tau \frac{R_\odot}{c} \approx 10^{11} \times 2.3 \text{ s} \approx 7000 \text{ yr}, \quad (16.3)$$

where for the last evaluation, it is handy to recall again that $1 \text{ yr} \approx \pi \times 10^7 \text{ s}$.

Once the photons reach the surface, they can escape the star and travel unimpeded through space, taking, for example, only a modest time $t_{earth} = \text{au}/c \approx 8 \text{ min}$ to cross the 1 au ($\approx 215R_\odot$) distance from the sun to the earth.

A stellar atmospheric surface thus marks a quite distinct boundary between the interior and free space. From deep within the interior, the stellar radiation field would appear nearly isotropic (same in all directions), with only a small asymmetry (of order $1/\tau$) between upward and downward photons. But near the surface, this radiation becomes distinctly anisotropic, emerging upward from the surface below, but with no radiation coming downward from empty space above.

16.2 Diffusion approximation at depth

This picture of photons undergoing a random walk through the stellar interior can be formalized in terms of a *diffusion* model for radiation transport in the interior. Appendix D discusses the transition from diffusion to free-streaming that occurs in the narrow region near the stellar surface, known as the “stellar atmosphere”. This is described by the *equation of radiative transfer*, given by eqn. (D.1), with eqn. (D.2) now defining the vertical optical depth $\tau(r)$ from a given radius r to an external *observer* at $r \rightarrow \infty$.

But in the deep interior layers within a star, i.e. with large optical depths $\tau \gg 1$, the trapping of the radiation makes the intensity I nearly isotropic and near the local Planck function B . Applying this to the derivative term in eqn. (D.1) and solving for I gives a “diffusion approximation” form for the intensity,

$$I(\mu, \tau) \approx B(\tau) + \mu \frac{dB}{d\tau}, \quad (16.4)$$

where we recall from §12.1 that μ is the cosine of the angle between the ray and the vertical direction, so that $\mu = +1$ is directly upward, and $\mu = -1$ is directly downward.

Since $dB/d\tau$ is of order B/τ , we can see that the second term is much smaller, by a factor $\sim 1/\tau \ll 1$, than the first, leading-order term. Recall that both the specific intensity I and the Planck function B have the same units as a surface brightness, i.e. energy/area/time/solid angle.

The local *net upward flux* F (energy/area/time) is computed by weighting the intensity by the direction cosine μ and then integrating over solid angle,

$$F \equiv \oint I \mu d\Omega = 2\pi \int_{-1}^{+1} \left(\mu B(\tau) + \mu^2 \frac{dB}{d\tau} \right) d\mu. \quad (16.5)$$

Since the leading order term with $B(\tau)$ is then odd over the range $-1 < \mu < 1$, it vanishes upon integration, giving

$$F \approx \frac{4\pi}{3} \frac{dB}{d\tau}. \quad (16.6)$$

Again recalling that $B = \sigma_{sb} T^4/\pi$, and noting the optical depth changes with radius r as $d\tau = -\kappa\rho dr$, we can alternatively write the flux as a function of the local temperature gradient,

$$F(r) = - \left[\frac{4\pi}{3\kappa\rho} \frac{\partial B}{\partial T} \right] \frac{dT}{dr} = - \left[\frac{16\sigma_{sb}}{3\kappa\rho} T^3 \right] \frac{dT}{dr}. \quad (16.7)$$

The terms in square bracket can be thought of as a *radiative conductivity*, which we note increases with the cube of the temperature T^3 , but depends inversely on opacity and density, $1/\kappa\rho$.

16.3 Atmospheric variation of temperature with optical depth

A star's luminosity L is generated in a very hot, dense central core. Outside this core, at any stellar envelope radius r , the local radiative flux scales as $F = L/4\pi r^2$, which near the stellar surface $r \lesssim R$ approaches the fixed surface value $F_* = L/4\pi R^2 \equiv \sigma_{sb} T_{\text{eff}}^4$, where the last equation recalls the definition of the stellar effective temperature T_{eff} . Since in such a surface layer F_* is independent of τ , eqn. (16.6) can be trivially integrated in this layer to give,

$$\frac{4\pi}{3} B(\tau) = F_* \tau + C = \sigma_{sb} T_{\text{eff}}^4 \tau + C, \quad (16.8)$$

where C is an integration constant. Recalling also from eqn. (5.1) that $\pi B = \sigma_{sb} T^4$, we can convert (16.8) into an explicit expression for the variation of temperature with optical depth,

$$T^4(\tau) = \frac{3}{4} T_{\text{eff}}^4 [\tau + 2/3], \quad (16.9)$$

wherein, in light of the result in §D.2, we have taken the integration constant such that $T(\tau = 2/3) = T_{\text{eff}}$.

Together with the equation (15.1) for hydrostatic equilibrium, equation (16.7) for radiative diffusion determines the fundamental structure of the stellar interior. The next section uses these to explain the underling physics behind the main-sequence, mass-luminosity relation $L \sim M^3$, which as discussed in §10 was found empirically from observations of binary systems with known parallactic distances (see fig. 10.4).

16.4 Questions and Exercises

Quick Question 1:

- At what optical depth τ does the local temperature T in a stellar atmosphere equal the stellar effective temperature T_{eff} ?
- At about what optical depth τ does the local temperature $T = 10T_{\text{eff}}$?

Quick Question 2:

- Near the Sun's surface where the temperature is at the effective temperature $T = T_* = T_{\text{eff}} \approx 5800 \text{ K}$, compute the scale height H (in km).
- Using the fact that the mean-free-path $\ell \approx H$ near this surface, compute the mass density ρ (in g/cm^3) assuming the opacity is equal to the electron scattering value given in §C.1, i.e. $\kappa_e = 0.34 \text{ cm}^2/\text{g}$.

17 Structure of Radiative vs. Convective Stellar Envelopes

17.1 $L \sim M^3$ relation for hydrostatic, radiative stellar envelopes

As discussed in part I (§10.4), observations of binary systems indicate that main sequence stars follow an empirical mass-luminosity relation $L \sim M^3$. The physical basis for this can be understood by considering the two basic relations of stellar structure, namely hydrostatic equilibrium and radiative diffusion, as given in eqns. (15.1) and (16.7) above.

As in the Virial scaling for internal temperature given in §15.3, we can use a single point evaluation of the hydrostatic pressure gradient to derive a scaling between interior temperature T , stellar radius R and mass M , and molecular weight μ ,

$$\begin{aligned}\frac{dP}{dr} &= -\rho \frac{GM_r}{r^2} \\ \rho \frac{T}{\mu R} &\sim \rho \frac{M}{R^2} \\ TR &\sim M\mu, \end{aligned} \tag{17.1}$$

Likewise, a single point evaluation of the temperature gradient in the radiative diffusion equation (16.7) gives

$$\begin{aligned}F &= -\frac{16\sigma_{sb}}{3\kappa\rho} T^3 \frac{dT}{dr} \\ \frac{L}{R^2} &\sim \frac{R^3}{\kappa M} \frac{T^4}{R} \\ L &\sim \frac{(RT)^4}{\kappa M} \\ L &\sim \frac{M^3 \mu^4}{\kappa}, \end{aligned} \tag{17.2}$$

where the last scaling uses the hydrostatic equilibrium scaling in (17.1) to derive the basic scaling law $L \sim M^3$, assuming a fixed molecular weight μ and stellar opacity κ .

Two remarkable aspects of this derivation are that: (1) the role of the stellar *radius cancels*; and (2) the resulting $M - L$ scaling does *not* depend on the details of the nuclear generation of the luminosity in the stellar core! Indeed, this scaling was understood from stellar structure analyses that were done (e.g. by Eddington, and Schwarzschild) in the 1920's, long before Hydrogen fusion was firmly established as a key energy source for the Sun and other main-sequence stars (e.g., by Hans Bethe ca. 1939).

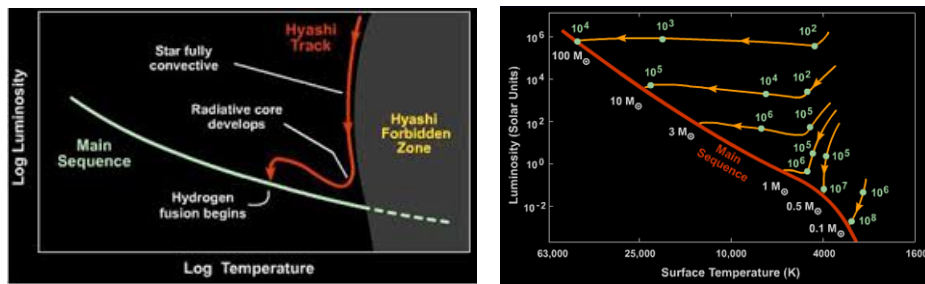


Figure 17.1 Illustration of the pre-main-sequence evolution of stars. The left panel shows how during the early stages of a collapsing proto-star, the interior is fully convective, causing it to evolve with decreasing luminosity at a nearly constant, relatively cool surface temperature, and so down the nearly vertical “Hayashi track” in the H-R diagram. The right panel shows the final approach to the main sequence for stars of various masses. For stars with a solar mass or above, the stellar interior becomes radiative, stopping the Hayashi track decline in luminosity. The stars then evolve horizontally and to the left on the H-R diagram, each with fixed luminosity but increasing temperature, till they reach their respective positions on the “zero-age-main-sequence” or ZAMS, when the core is hot enough to ignite H-fusion.

17.2 Horizontal-track Kelvin-Helmholtz contraction to the main sequence

In fact, this simple $L \sim M^3$ scaling even applies to the final stages of *pre*-main-sequence evolution, when the core is not yet hot enough to start nuclear burning, but the envelope has become hot enough for radiative diffusion to dominate the transport of energy generated by the star’s gravitational contraction. As the radius decreases over the Kelvin-Helmholtz timescale t_{KH} of this contraction, the surface temperature increases in a way that keeps the luminosity nearly constant. Figure 17.1 illustrates that, on the H-R diagram, a late-phase pre-main-sequence star with a mass near the Sun or higher thus evolves along a *horizontal track* from right to left, stopping when it reaches the main sequence; this is where the core temperature is now high enough for H-fusion to take over in supplying the energy for the stellar luminosity, without any need for further contraction. As discussed in §18, for a given mass, a star’s radius on the main

sequence is just the value for which the interior temperature, as set by the Virial theorem, is sufficiently high to allow this H-fusion in the core.

17.3 Convective instability and energy transport

In practice, the transport of energy from the stellar interior toward the surface sometimes occurs through *convection* instead of radiative diffusion; this has important consequence for stellar structure and thus for the scaling of luminosity.

Convection refers to the overturning motions of the gas, much like the bubbling of boiling water on a stove. Stars become unstable to forming convection whenever the processes controlling the temperature make its spatial gradient too steep. This can occur in the nuclear burning core of massive stars, for which the specific mechanism for Hydrogen fusion, called the “CNO” cycle, gives the nuclear burning rate a steep dependence on temperature (§18). The resulting steep temperature gradient makes the cores of such stars strongly convective.

Steep gradients, and their associated convection, can also occur in outer regions of cooler, lower-mass stars, where the cooler temperature induces recombination of ionized H or He. The bound-free absorption by this neutral Hydrogen significantly increases the local stellar opacity κ . For a fixed stellar flux $F = L/4\pi r^2$ of stellar luminosity L that needs to be transported through an interior radius r , the radiative diffusion eqn. (16.7) shows that the required radiative temperature gradient increases with such increased opacity,

$$\left| \frac{dT}{dr} \right|_{rad} = \frac{3\kappa\rho F}{16\sigma_{sb}T^3} \sim \kappa \quad (17.3)$$

If this gradient becomes too steep, then, as illustrated in figure 17.2, a small element of gas that is displaced slightly upward becomes less dense than its surroundings, giving it a buoyancy that causes it to rise higher still. A key assumption is that this dynamical rise of the fluid occurs much more rapidly than the rate for energy to diffuse into or out of the gas element. Processes that occur without any such energy exchange with the surroundings are called “*adiabatic*”, with a fixed (power-law) relation of pressure with density or temperature. In a hydrostatic medium with a set pressure gradient, this implies a fixed adiabatic temperature gradient $(dT/dr)_{ad}$.

Starting from an initial radius r_0 with equal density and temperature inside and outside some chosen fluid element (i.e., $\rho'_0 = \rho$, $T'_0 = T_0$), let us determine the density ρ'_1 of that element after it is adiabatically displaced to a slightly higher radius $r_1 = r_0 + \delta r$, where the ambient density is ρ_1 . Since dynamical balance requires the element and its surrounding to still have equal pressure after the displacement (i.e., $P'_1 = P_1$), we have by the perfect gas law that $\rho'_1 T'_1 = \rho_1 T_1$. If this upward displacement $\delta r > 0$ makes the element buoyant, with lower density ρ'_1 than that of its surroundings ρ_1 , then using this constant pressure condition, we can derive the condition for the temperature gradient required for

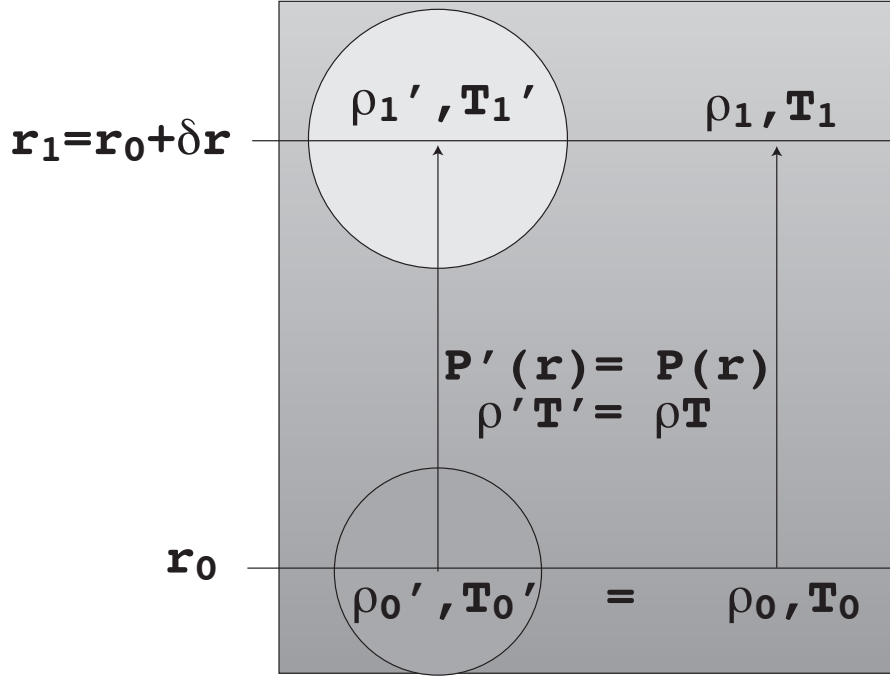


Figure 17.2 Illustration of upward displacement of a spherical fluid element in test for convective instability, which occurs when the displaced element has a lower density ρ'_1 than that of its surroundings, ρ_1 . Since the pressure must remain equal inside and outside the element, this requires the element to have a higher temperature, $T'_1 > T_1$. Since the overall temperature gradient is negative, convection thus occurs whenever the magnitude of the atmospheric temperature gradient is steeper than the adiabatic gradient that applies for the adiabatically displaced element, i.e., $|dT/dr| > |dT/dr|_{ad}$.

the associated convective instability,

$$\frac{T_1}{T'_1} = \frac{\rho'_1}{\rho_1} < 1 \quad ; \quad \text{Convective instability}$$

$$T_0 + \delta r (dT/dr)_{rad} = T_1 < T'_1 = T_0 + \delta r (dT/dr)_{ad}$$

$$\left| \frac{dT}{dr} \right|_{rad} > \left| \frac{dT}{dr} \right|_{ad}, \quad (17.4)$$

where since both temperature gradients are negative, the condition in terms of absolute value requires a reversal of the inequality.

We thus see that convection will ensue whenever the magnitude of the radiative temperature gradient exceeds that of the adiabatic temperature gradient.

Convection is an inherently complex, 3D dynamical process that generally

requires elaborate computer simulations to model accurately. A heuristic, semi-analytic model called “mixing length theory” has been extensively developed, but it has serious limitations, especially near the stellar surface, where the lower density and temperature can make convective transport quite inefficient. By contrast, in the dense and hot stellar interior, once convection sets in, it is so efficient at transporting energy that it keeps the local temperature gradient very close to the adiabatic value above which it is triggered.

One can thus quite generally just presume that the temperature gradient in interior convection regions is at the adiabatic value.

17.4 Fully convective stars – the Hayashi track for proto-stellar contraction

In hot stars with $T > 10,000$ K, Hydrogen remains fully ionized even to the surface; since there then is no recombination zone to increase the opacity and trigger convection, the energy transport in their stellar envelopes is by radiative diffusion. In moderately cooler stars like the Sun (with $T_{\odot} \approx 6000$ K), Hydrogen recombination in a zone just somewhat below the surface induces convection, which thus provides the final transport of energy toward the surface; but since the deeper interior remains ionized and thus non-convective, the general scaling laws derived assuming radiative transport still roughly apply for such solar-type stars.

However, in much cooler stars, with surface temperatures $T \approx 3500 - 4000$ K, the Hydrogen recombination extends deeper into the interior; this and other factors keep the opacity high enough to make the entire star convectively unstable right down to the stellar core. Because convection is so much more efficient than radiative diffusion, it can readily bring to the surface any energy generated in the interior – whether produced by gravitational contraction of the envelope, or by nuclear fusion in the core. As such, *fully convective stars* can have luminosities that greatly exceed the value implied by the $L \sim M^3$ scaling law derived in §17.1 (see eqn. 17.2) for stars with radiative envelopes. As discussed in §19, this is a key factor in the high luminosity of cool giant stars that form in the post-main-sequence phases after the exhaustion of Hydrogen fuel in the core.

But it also helps explain the high luminosity of the very cool, early stage of *pre-main-sequence* evolution, when gravitational contraction of a large *proto-stellar cloud* is providing the energy to make the cloud shine as a *proto-star*. Once the internal pressure generated is sufficient to establish hydro-static equilibrium, its interior becomes fully convective, forcing the proto-star to have this characteristic surface temperature around $T \approx 3500 - 4000$ K.

At early stages the proto-star’s radius is very large, meaning it has a very large luminosity $L = \sigma_{sb} T^4 4\pi R^2$. As it contracts, it stays at this temperature, but the declining radius means a declining luminosity. As illustrated in figure 17.1, during this early phase of gravitational contraction, the proto-star thus

evolves down a nearly vertical line in the H-R diagram, dubbed the “Hayashi” track, after the Japanese scientist who first discovered its significance.

Once the radius reaches a level at which the luminosity is near the value predicted by the $L \sim M^3$ law, the interior switches from convective to radiative, and so the final contraction to the main sequence makes a sharp turn to a horizontal track (sometimes called the “Heneyey” track) with nearly constant luminosity but decreasing surface temperature. The luminosity of this track is set by the stellar mass, according to the $L \sim M^3$ law derived for stars with interior energy transport by radiative diffusion. The contraction is halted when the core reaches a temperature (derived in the section 18; see equation 18.4) for H-fusion, which then stably supplies the luminosity for the main-sequence lifetime.

As detailed in §19, once the star runs out of Hydrogen fuel in its core, its *post*-main-sequence evolution effectively traces backwards along nearly the same track followed during this *pre*-main-sequence, ultimately leading to the cool, red giant stars seen in the upper right of the H-R diagram.

18 Hydrogen Fusion and the Mass Range of Stars

The timescale analyses in part I (§8) show that nuclear fusion of Hydrogen into Helium provides a long-lasting energy source that we can associate with main sequence stars in the H-R diagram (§6.3). But what are the requirements for such fusion to occur in the stellar core? And how is this to be related to the luminosity vs. surface temperature scaling for main sequence stars in the HR diagram? In particular, how might this determine the relation between mass and radius? Finally, what does it imply about the lower mass limit for stars to undergo Hydrogen fusion?

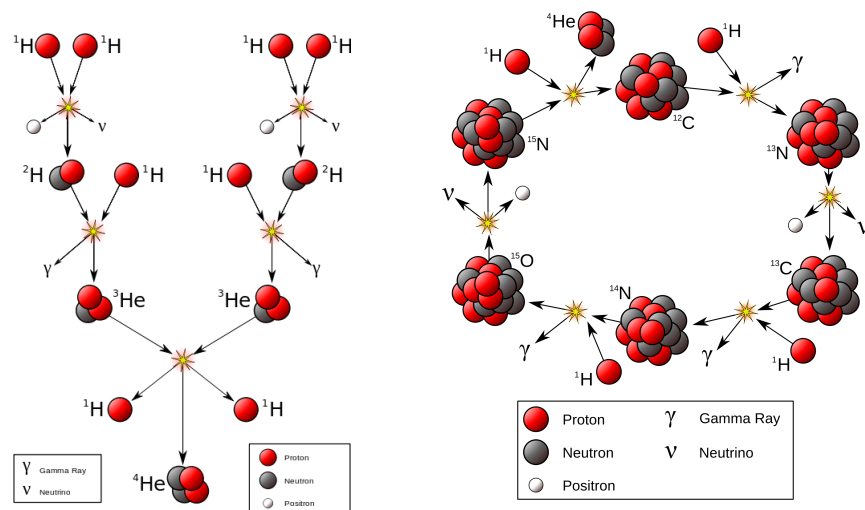
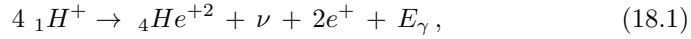


Figure 18.1 The two distinct channels for hydrogen fusion in stellar cores. For the Sun and other low-mass stars, this occurs by the direct proton-proton (PP) chain (left). For high-mass stars, it occurs via the CNO cycle (right), in which Carbon, Nitrogen and Oxygen nuclei serve as catalysts for the overall fusion of Hydrogen into Helium. The higher charge of CNO nuclei requires higher proton energy to overcome the higher electrical repulsion. This makes CNO burning very sensitive to temperature, and so dominant in the hotter cores of higher-mass stars. Credit: Borb.

18.1 Core temperature for H-fusion

Figure 18.1 illustrates that there are two distinct channels for fusing hydrogen into helium in stellar cores: the direct proton-proton (PP) chain on the left; and the CNO cycle on the right. The latter turns out to be dominant in more massive stars; their higher core temperatures makes it possible for protons to overcome the higher electrical repulsion of the higher charges of CNO nuclei, allowing these then to become effective *catalysts* for a net fusion of Hydrogen into Helium.

But in the Sun and other low-mass stars, the core temperatures are only sufficient for direct PP-chain fusion. The left panel of figure 18.1 illustrates the most important of the detailed reaction channels, but the overall result is simply



where ν represents a weakly interacting neutrino (which simply escapes the star). The $2e^+$ represents two positively charged “anti-electrons”, or *positrons*, which quickly annihilate with ordinary electrons, releasing $\sim 2 \times 2 \times \frac{1}{2} \approx 2\text{ MeV}$ of energy. The rest of the net $\sim 4 \times 7\text{ MeV}$ in energy, representing the mass-energy difference between $4H$ vs. one He , is released as high-energy photons (γ -rays) of energy E_γ .

The essential requirement for such PP fusion is that the thermal kinetic energy kT of the protons overcome the mutual repulsion of their positive charge $+e$, to bring the protons to a close separation at which the strong nuclear (attractive) force is able take over, and bind the protons together. For a given temperature T , the minimum separation b for two protons colliding head-on comes from setting this thermal kinetic energy equal to the electrostatic repulsion energy,

$$kT = \frac{e^2}{b}. \quad (18.2)$$

In particular, if we were to require that this minimum separation be equal to the size of a Helium nucleus, i.e. $b \approx 1\text{ fm} = 10^{-15}\text{ m}$, then from eqn. (18.2) we would infer that the required temperature is quite extreme, $T \approx 1.7 \times 10^{10}\text{ K}$!

Comparison with the virial scaling (15.9) shows this is more than a *thousand* times the characteristic virial temperature for the solar interior, $T_{int} \approx 13\text{ MK}$. As such, the closest distance b between protons in the interior core of the Sun is actually more than a thousand times the size of the Helium nucleus, which is thus well outside the scale for operation of the strong nuclear force that keeps the nucleus bound.

The reason that nuclear fusion can nonetheless proceed at such a relatively modest temperature stems again from the uncertainty principle of modern quantum physics. Namely, a proton with thermal energy $m_p v_{th}^2/2 = kT$ has an associated momentum $p = m_p v_{th} = \sqrt{2m_p kT}$. Within quantum mechanics, it thus has an associated ‘fuzziness’ in position, characterized by its De Broglie wavelength $\lambda \equiv h/p$, where h is Planck’s constant. If $\lambda \gtrsim b$, then there is a good probability that this waviness of protons will allow them to ‘tunnel’ through the

electrostatic repulsion barrier between them, and so find themselves within a nuclear distance at which the strong attractive nuclear force can bind them. Setting $b = \lambda = h/(m_p v_{th})$ in eqn. (18.2), we can thus obtain an explicit expression for the proton thermal speed needed for nuclear fusion of Hydrogen¹,

$$v_{th,nuc} = \frac{2e^2}{h} = 690 \text{ km/s}. \quad (18.3)$$

Two remarkable aspects of eqn. (18.3) are: (1) this thermal speed for H-fusion depends *only* on the fundamental physics constants e and h , and (2) its numerical value is very nearly equal to the surface escape speed from the Sun, $v_{esc} = \sqrt{2GM_\odot/R_\odot} = 618 \text{ km/s}$. Recalling the virial scaling (15.9) that says the thermal energy in the stellar interior is comparable to the gravitational binding energy, this means that given the solar mass M_\odot the Sun has adjusted to just the radius needed for the gravitational binding to give an interior temperature that is hot enough for Hydrogen fusion. For mean molecular weight $\bar{\mu} \approx 0.6m_p$, the mean thermal speed (18.3) implies a core temperature

$$T_{nuc} = \frac{\bar{\mu} v_{th,nuc}^2}{2k} = 1.2 \frac{m_p e^4}{k h^2} \approx 17 \text{ MK}, \quad (18.4)$$

which now is quite comparable to the interior temperature $T_{int,vir} \approx 13 \text{ MK}$ obtained by applying the virial scaling (15.9) to the Sun.

18.2 Main sequence scalings for radius-mass and luminosity-temperature

If we were to naively apply these same scalings to stars with different masses, then it would suggest all stars along the main sequence should have the same, *solar ratio of mass to radius*, and thus that the radius should increase linearly with mass, $R \sim M$.

In practice, the radius-mass relation for main-sequence stars is somewhat sub-linear,

$$R \sim M^{0.7}. \quad (18.5)$$

This can be understood by considering that the much higher luminosity of more massive stars, scaling as $L \sim M^3$, means that the core – within which the total fuel available scales just linearly with stellar mass M – must have more vigorous nuclear burning². The higher core temperature to drive such more vigorous H-

¹ I am indebted to Prof. D. Mullan for pointing out to me this remarkably simple scaling.

² Indeed, as already noted, in massive stars the standard, direct proton-proton fusion is augmented by a process called the CNO cycle, in which CNO elements act as a catalyst for H-fusion. Attaching protons to such more highly charged CNO nuclei requires a higher core temperature to overcome the stronger electrical repulsion, and this indeed obtains in such massive stars.

fusion then requires by the Virial theorem that the mass to radius ratio of such stars must be somewhat higher than for lower mass stars like the Sun.

Combining such a sublinear radius-mass scaling $R \sim M^{0.7}$ with the mass-luminosity scaling $L \sim M^3$ (eqn. 17.2) and the Stefan-Boltzmann relation $L \sim T^4 R^2$, we infer that luminosity should be a quite steep function of surface temperature along the main sequence, viz. $L \sim T^8$. While observed HR diagrams (like that plotted for nearby stars in part I) show the main sequence to have some complex curvature structure, a straight line with $\log L \sim 8 \log T$ does give a rough overall fit, thus providing general support for these simple scaling arguments.

18.3 Lower mass limit for hydrogen fusion: Brown Dwarf stars

These nuclear burning scalings can also be used to estimate a minimum stellar mass for Hydrogen fusion. Stars with mass below this minimum are known as Brown Dwarfs. A key new feature of these stars is that their cores become “*electron degenerate*”, and so no longer follow the simple virial scalings derived above for stars in which the pressure is set by the ideal gas law. Electron degeneracy occurs when the electron number density n_e becomes comparable to cube of the electron De Broglie wavenumber $k_e \equiv 2\pi/\lambda_e \equiv 1/\bar{\lambda}_e$,

$$n_e \approx k_e^3 = \frac{1}{\bar{\lambda}_e^3}, \quad (18.6)$$

with the electron thermal De Broglie (reduced) wavelength,

$$\bar{\lambda}_e = \frac{\hbar}{p_e} = \frac{\hbar}{\sqrt{2m_e kT}}, \quad (18.7)$$

where $\hbar \equiv h/2\pi$, and the latter equality casts the electron thermal momentum p_e in terms the temperature T and electron mass m_e . Assuming a constant density $\rho = M/(4\pi R^3/3) \approx m_p n_e$, we can combine (18.6) and (18.7) with the nuclear temperature (18.4) and the Virial relation (15.9) to obtain a relation for the stellar mass at which a nuclear burning core should become electron degenerate,

$$M_{\min, \text{nuc}} = \frac{\sqrt{3/2}}{4\pi^2} \left(\frac{m_p}{m_e} \right)^{3/4} \frac{e^3}{G^{3/2} m_p^2} \approx \boxed{0.1 M_\odot}. \quad (18.8)$$

Stars with a mass below this minimum should not be able to ignite H-fusion, because electron degeneracy prevents their cores from contracting to a small enough size to reach the ~ 17 MK temperature (see eqn. 18.4) required for fusion. In practice, more elaborate computations indicate such Brown Dwarf stars have a limiting mass $M_{BD} \lesssim 0.08 M_\odot$, just slightly below the simple estimate given in (18.8).

Note that, although this minimum mass for H-fusion is limited by electron degeneracy, the actual value is *independent* of Planck’s constant h ! Essentially,

the role of h in the tunneling effect for H-fusion *cancels* its role in electron degeneracy.

18.4 Upper mass limit for stars: the Eddington Limit

Let's next consider what sets the *upper* mass limit for observed stars. This is not linked to nuclear burning or degeneracy, but stems from the strong $L \sim M^3$ scaling of luminosity with mass, which, as noted in §17.1, follows from the hydrostatic support and radiative diffusion of the stellar envelope.

In addition to its important general role as a carrier of energy, radiation also has an associated momentum. For a photon of energy $E = h\nu$, the associated momentum is set by its energy divided by the speed of light, $p = h\nu/c$. The trapping of radiative energy within a star thus inevitably involves a trapping of its associated momentum, leading to an outward radiative force, or for a given mass, an outward *radiative acceleration* g_{rad} , that can compete with the star's gravitational acceleration g . For a local radiative energy flux F (energy/time/area), the associated momentum flux (force/area, or pressure) is just F/c . The material acceleration resulting from absorbing this radiation depends on the effective cross sectional area σ for absorption, divided by the associated material mass m , as characterized by the opacity κ ,

$$g_{\text{rad}} = \frac{\sigma}{m} \frac{F}{c} = \frac{\kappa F}{c}. \quad (18.9)$$

For a star of luminosity L , the radiative flux at some radial distance r is just $F = L/4\pi r^2$. This gives the radiative acceleration the same inverse-square radial decline as the stellar gravity, $g = GM/r^2$, meaning that it acts as a kind of “anti-gravity”.

Sir Arthur Eddington first noted that, even for a minimal case in which the opacity just comes from free-electron scattering, $\kappa = \kappa_e = 0.2(1 + X) \approx 0.34 \text{ cm}^2 \text{ g}^{-1}$ (with the numerical value for standard (solar) Hydrogen mass fraction $X \approx 0.7$; see Appendix C), there is a limiting luminosity, now known as the “*Eddington luminosity*”, for which the radiative acceleration $g_{\text{rad}} = g$ would completely *cancel* the stellar gravity,

$$L_{\text{Edd}} = \frac{4\pi GMc}{\kappa_e} = 3.8 \times 10^4 L_{\odot} \frac{M}{M_{\odot}}. \quad (18.10)$$

Any star with $L > L_{\text{Edd}}$ is said to exceed the “*Eddington limit*”, since even the radiative acceleration from just scattering by free electrons would impart a force that exceeds the stellar gravity, thus implying that the star would no longer be gravitationally bound!

For main sequence stars that follow the $L \sim M^3$ scaling, setting $L = L_{\text{Edd}}$ yields an estimate for an upper mass limit at which the star would reach this Eddington limit,

$$\boxed{M_{\text{max,Edd}} \approx 195 M_{\odot}}, \quad (18.11)$$

where $195 \approx \sqrt{3.8 \times 10^4}$. This agrees quite well with modern empirical estimates for the most massive observed stars, which are in the range 150-300 M_\odot .

Actually, as stars approach this Eddington limit, the radiation pressure alters the hydrostatic structure of the envelope, causing the mass-luminosity relation to weaken toward a linear scaling, $L \sim M$, and so allowing in principle for stars with even with mass $M > M_{max,Edd}$ to remain bound. In practice, such stars are subject to “photon bubble” instabilities, much as occurs whenever a heavy fluid (in this case the stellar gas) is supported by a lighter one (here the radiation). Very massive stars near this Eddington limit thus tend to be highly variable, often with episodes of large ejection of mass that effectively keeps the stellar mass near or below the $M_{max,Edd} \approx 195M_\odot$ limit.

Exercise 1:

- Assume a power-law radius-mass scaling $R \sim M^a$ for stars on the main-sequence. Show that there is an associated power-law relation $L \sim T^b$ between luminosity L and surface temperature T .
- Give a formula for the power-index b in terms of the index a .
- Compute the values for b for cases with $a = 0.5, 0.7$ and 1 .
- What value of a would give the $L \sim T^8$ quoted in the text.

19 Post-Main-Sequence Evolution: Low-Mass Stars

As a star ages, more and more of the Hydrogen in its core becomes consumed by fusion into Helium. Once this core Hydrogen is used up, how does the star react and adjust? Without the H-fusion to supply its luminosity, one might think that perhaps the star would simply shrink, cool and dim, and so die out, much as a candle when all its wax is used up.

Instead it turns out that stars at this *post-main-sequence* stage of life actually start to *expand*, at first keeping roughly the same luminosity and so becoming cooler at the surface, but eventually becoming much brighter giant or supergiant stars, shining with a luminosity that can be thousands or even tens of thousands that of their core-H-burning main sequence.

Figure 19.1 illustrates the post-MS evolution for the Sun (left) and for stars with mass up to $10 M_{\odot}$ (right).

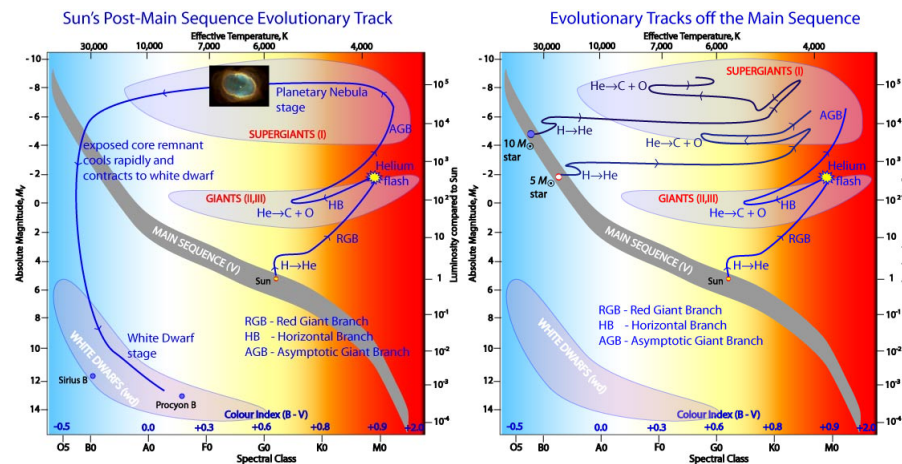


Figure 19.1 Schematic H-R diagrams to show the post-main-sequence evolution for a solar-mass star (left), and for stars with $M = 1, 5$, and $10 M_{\odot}$ (right).

As summarized in figure 19.2, the evolution and final states of stars depends on the stellar mass, with distinct difference for stars with initial masses below vs. above about $8 M_{\odot}$. The remainder of this section focuses out initial attention

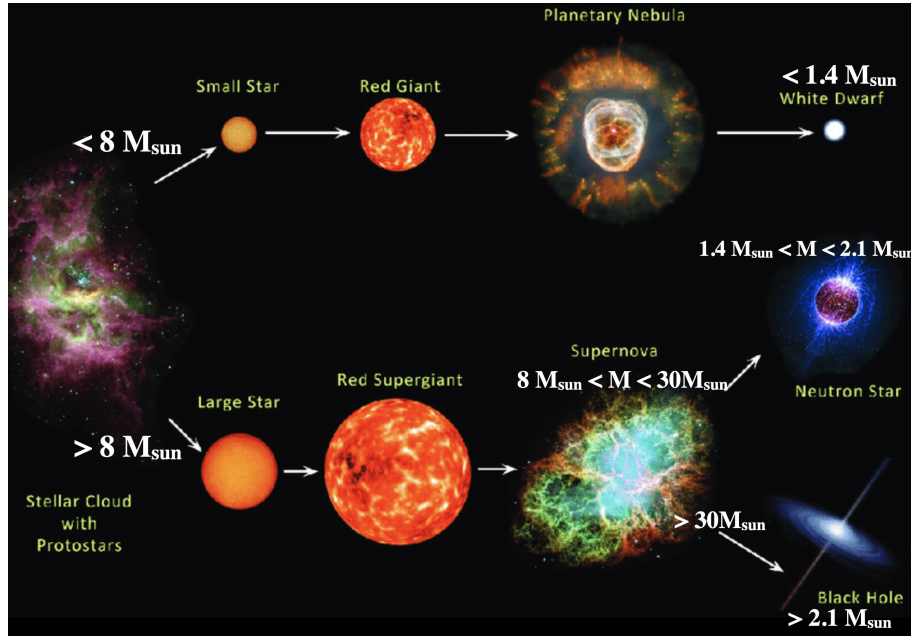


Figure 19.2 Distinct evolution and final states for stars with initial masses above and below $8M_{\odot}$.

on solar-type stars with $M \lesssim 8M_{\odot}$. The evolution and final states of high-mass stars is discussed in the next section (§20).

19.1 Core-Hydrogen burning and evolution to the Red Giant branch

The apparently counterintuitive *post*-main-sequence adjustment of stars can actually be understood through the same basic principals used to understand their initial, *pre*-main-sequence evolution. When the core runs out of Hydrogen fuel, the lack of energy generation does indeed cause the core itself to contract. But the result is to make this core even denser and hotter. Then, much as the hot coals at the heart of a wood fire help burn the wood fuel around it much faster, the higher temperature of a contracted stellar core actually makes the overlying shell of Hydrogen fuel around the core burn even more vigorously!

Now, unlike during the main sequence – when there is an essential regulation or compatibility between the luminosity generated in the core and the luminosity that the radiative envelope is able to transport to the stellar surface –, this shell-burning core is actually *over-luminous* relative to the envelope luminosity that is set by the $L \sim M^3$ scaling law. As such, instead of emitting this core luminosity as surface radiation, the excess energy acts to *expand* the star, in effect doing work against gravity to *reverse* the Kelvin-Helmholtz contraction that occurred

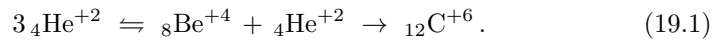
during the star's *pre*-main-sequence evolution. Initially, the radiative envelope keeps the luminosity fixed so that, as the star expands, the surface temperature again declines, with the star thus again evolving horizontally on the H-R diagram, this time from left to right.

But as the surface temperature approaches the limiting value $T \approx 3500 - 4000$ K, the envelope again becomes more and more convective, which thus now allows this full high-luminosity of the H-shell-burning core to be transported to the surface. The star's luminosity thus increases, with now the temperature staying nearly constant at the cool value for the Hayashi limit. In the H-R diagram, the star essentially climbs back up the Hayashi track, eventually reaching the region of the cool, red giants in the upper right of the H-R diagram.

The above describes a general process for all stars, but the specifics depend on the stellar mass. For masses less than the Sun, the main sequence temperature is already quite close to the cool limit, so evolution can proceed almost directly vertically up the Hayashi track. For masses much greater than the Sun, the luminosity and temperature on the main sequence are both much higher, and so the horizontal evolutionary phase is more sustained. And since the luminosity is already very high, these stars become red *supergiants* without ever having to reach or climb the Hayashi track.

19.2 Helium flash and core-Helium burning on the Horizontal Branch

This Hydrogen-shell burning also has the effect of increasing further the temperature of the stellar core, and eventually this reaches a level where the fusion of the Helium itself becomes possible, through what's known as the "triple- α process"¹,



The direct fusion of two ${}_4\text{He}^{+2}$ nuclei initially make an unstable nucleus of Beryllium (${}_8\text{Be}^{+4}$), which usually just decays back into Helium. But if the density and temperature are sufficiently high, then during the brief lifetime of the unstable Beryllium nucleus, another Helium can fuse with it to make a very stable Carbon nucleus ${}_{12}\text{C}^{+6}$. Since the final step of fusing ${}_4\text{He}^{+2}$ and ${}_8\text{Be}^{+4}$ involves overcoming an electrostatic repulsion that is $2 \times 4 = 8$ times higher than for proton-proton (p-p) fusion of Hydrogen, He-fusion requires a much higher core temperature, $T_{\text{He}} \approx 8 \times 15 \text{ MK} \approx 120 \text{ MK}$.

In stars with more than a few solar masses, this ignition of the Helium in the core occurs gradually, since the higher core temperature from the addition of

¹ Since Helium nuclei are sometimes referred as " α -particles". In this formula, the left subscript denotes the atomic mass (number of proton and neutrons), while the right superscript denotes the nuclear charge (number of protons).

He-burning increases the gas pressure, making the core tend to expand in a way that regulates the burning rate.

In contrast, for the Sun and other stars with masses $M < 2M_{\odot}$, the number density of electrons n_e in the helium core is so high² that their core becomes *electron degenerate*. As discussed in §18.3 for the Brown dwarf stars that define the lower mass limit for H-burning, electron degeneracy occurs when the mean distance between electrons $\sim n_e^{-1/3}$ becomes comparable to the DeBroglie wavelength $\bar{\lambda}_e = \hbar/p_e$. The properties of such degeneracy are discussed further in §19.4 on the degenerate white-dwarf end states of solar-type stars.

But in the present context a key point is that the response to any heat addition is quite different than for an ideal gas. By the virial theorem for a gravitationally bound ideal gas, the added heating from any increase in nuclear burning leads to an expansion that cools the gas, thus reducing the burning and so keeping it stable. In contrast, for a degenerate gas, the expansion from adding heat actually makes the temperature increase even further. Thus, once the evolutionary increase in core temperature reaches a level that ignites fusion of Helium into Carbon, the degenerate nature of the gas leads to a *Helium flash*, in which a substantial fraction of the core of Helium is fused into Carbon over a very short timescale.

This flash marks the “tip” of the Red Giant Branch (RGB) in the H-R diagram; but somewhat surprisingly, the sudden addition of energy is largely absorbed by the expansion of the core and the overlying stellar envelope. Since the expanded core is no longer very degenerate, the star thus simply settles down to a more quiescent, stable phase of He-burning. The expanded core also means the shell burning of H actually declines, causing the luminosity to decrease from the tip of the Red Giant branch, where the He flash occurs, to a somewhat hotter, dimmer region known as the “Horizontal Branch” in the H-R diagram.

This Horizontal Branch (HB) can be loosely thought of as the He-burning analog of the H-burning Main Sequence (MS), but a key difference is that it lasts a much shorter time, typically only 10 to 100 *million years*, much less than the many billion years for a solar mass star on the MS. This is partly because the luminosity for HB stars is so much higher than for a similar mass on the MS, implying a much higher burn rate of fuel. But another factor is that the energy yield per-unit-mass, ϵ , for He-fusion to Carbon is about a tenth of that for H-fusion to Helium, viz. about $\epsilon_{\text{He}} \approx 0.06\%$ vs. the $\epsilon_{\text{H}} \approx 0.7\%$ for H-burning (see figure 20.1). With the lower energy produced, and the higher rate of energy lost in luminosity, the lifetime is accordingly shorter.

² Recall that on the main sequence the radii of stars is (very) roughly proportional to their mass, $R \sim M$. But since density scales as $\rho \sim M/R^3$, the *density* of low mass stars tends generally to be higher than in high-mass stars, roughly scaling as $\rho \sim 1/M^2$. This overall scaling of average stellar density also applies to the relative densities of stellar cores, and so helps explain why the cores of low-mass stars tend to become electron degenerate, while those of higher mass stars do not.

19.3 Asymptotic Giant Branch to Planetary Nebula to White Dwarf

Once the core runs out of Helium, He-burning also shifts to an inner shell around the core, which itself is still surrounded by a outer shell of more vigorous H-burning. This again tends to increase the core luminosity, but now since the star is cool and thus mostly convective, this energy is mostly transported to the surface with only a modest further expansion of the stellar radius. This causes the star to again climb in luminosity along what's called the "*Asymptotic Giant Branch*" (AGB), which parallels the Hayashi track at just a somewhat hotter surface temperature.

In the Sun and stars of somewhat higher mass, up to $M \lesssim 8M_{\odot}$, there can be further ignition of the Carbon to fuse with Helium to form Oxygen. But further synthesis up the periodic table requires overcoming the greater electrical repulsion of more highly charged nuclei. This in turn requires a temperature higher than occurs in the cores of lower mass stars, for which the onset of electron degeneracy prevents contraction to a denser, hotter core. Further core burning thus ceases, leaving the core as an inert, degenerate ball of C and O, with final mass on order of $1 M_{\odot}$, with the remaining mass contained in the surrounding envelope of mostly Hydrogen.

But such AGB stars tend also to be pulsationally unstable, and because of the very low surface gravity, such pulsations can over time actually *eject the entire stellar envelope*. This forms a circumstellar *nebula* that is heated and ionized by the very hot remnant core. As seen in the left panel of figure 20.5, the resulting circular nebular emission glow somewhat resembles the visible disk of a planet, so these are called "*planetary nebulae*", though they really have nothing much to do with actual planets. After a few thousand years, the planetary nebula dissipates, leaving behind just the degenerate remnant core, a white dwarf star.

19.4 White Dwarf stars

The electron-degenerate nature of white dwarf stars endows them with some rather peculiar, even extreme properties. As noted, they typically consist of roughly a solar mass of C and O, but have a radius comparable to that of the Earth, $R_e \approx R_{\odot}/100$. This small radius makes them very dense, with $\rho_{\text{wd}} \approx 10^6 \bar{\rho}_{\odot} \approx 10^6 \text{ g/cm}^3$, i.e. about a million times (!) the density of water, and so a million times the density of normal main-sequence stars like the Sun. It also gives them very strong surface gravity, with $g_{\text{wd}} \approx 10^4 g_{\odot} \approx 10^6 \text{ m/s}^2$, or about 100,000 times Earth's gravity!

As noted in §18.3 for the Brown dwarf stars that define the lower mass limit for H-burning, a gas becomes electron degenerate when the electron number density n_e becomes so high that the mean distance between electrons becomes

comparable to their reduced De Broglie wavelength,

$$n_e^{-1/3} \approx \bar{\lambda} \equiv \frac{\hbar}{p_e} = \frac{\hbar}{m_e v_e}, \quad (19.2)$$

where the electron thermal momentum p_e equals the product of its mass m_e and thermal speed v_e , and $\hbar \equiv h/2\pi$ is the reduced Planck constant. The associated electron pressure is

$$P_e = n_e v_e p_e = n_e^{5/3} \frac{\hbar^2}{m_e} = \left(\frac{\rho Z}{A m_p} \right)^{5/3} \frac{\hbar^2}{m_e}, \quad (19.3)$$

where the last equality uses the relation between electron density and mass density, $\rho = n_e A m_p / Z$, with Z and $A m_p$ the average nuclear charge and atomic mass. For example, for a Carbon white dwarf, the atomic number $Z = 6$ gives the number of free (ionized) electrons needed to balance the $+Z$ charge of the Carbon nucleus, while the atomic weight $A m_p = 12 m_p$ gives the associated mass from the C atoms. The hydrostatic equilibrium (cf. eqn. 15.1) for pressure gradient support against gravity then requires for a white dwarf star with mass M_{wd} and radius R_{wd} ,

$$\frac{P_e}{R_{\text{wd}}} \approx \rho \frac{G M_{\text{wd}}}{R_{\text{wd}}^2}. \quad (19.4)$$

Using the density scaling $\rho \sim M_{\text{wd}}/R_{\text{wd}}^3$, we can combine (19.3) and (19.4) to solve for a relation between the white-dwarf radius and its mass,

$$R_{\text{wd}} = \frac{1}{G M_{\text{wd}}^{1/3}} \frac{\hbar^2}{m_e} \left(\frac{Z}{A m_p} \right)^{5/3} \approx \boxed{0.01 R_\odot \left(\frac{M_\odot}{M_{\text{wd}}} \right)^{1/3}}, \quad (19.5)$$

where the approximate evaluation uses the fact that for both C and O the ratio $Z/A = 1/2$. For a typical mass of order the solar mass, we see that a white dwarf is very compact, comparable to the radius of the Earth, $R_e \approx 0.01 R_\odot$. But note that this radius actually *decreases* with increasing mass.

19.5 Chandasekhar limit for white-dwarf mass: $M < 1.4M_\odot$

This fact that white-dwarf radii decrease with higher mass means that, to provide the higher pressure to support the stronger gravity, the electron speed v_e must strongly increase with mass. Indeed, at some point this speed approaches the speed of light, $v_e \approx c$, implying that the associated electron pressure now takes the scaling (cf. eqn. 19.3),

$$P_e = n_e c p_e = n_e^{4/3} \hbar c = \left(\frac{\rho Z}{A m_p} \right)^{4/3} \hbar c. \quad (19.6)$$

Applying this in the hydrostatic relation (19.4), we now find that the radius R *cancels*! Instead we can solve for a *upper limit* for a white dwarf’s mass,

$$M_{\text{wd}} \leq M_{\text{ch}} = \sqrt{3\pi} \left(\frac{\hbar c}{G} \right)^{3/2} \left(\frac{Z}{Am_{\text{p}}} \right)^2 \approx \boxed{1.4M_{\odot}}, \quad (19.7)$$

where the subscript refers to “Chandrasekhar”, the astrophysicist who first derived this mass limit, and the proportionality factor $\sqrt{3\pi}$ comes from a detailed calculation beyond the scope of the discussion here.

As discussed in later sections (§31.1), when accretion of matter from a binary companion puts a white dwarf over this limit, it triggers an enormous “white-dwarf supernova” explosion, with a large, relatively well-defined peak luminosity, $L \approx 10^{10} L_{\odot}$. This provides a very bright standard candle that can be used to determine distances as far as a Gpc, giving a key way to calibrate the expansion rate of the universe.

But in the present context, this limit means that sufficiently massive stars with cores above this mass cannot end their lives as a white dwarf. Instead, they end as violent “core-collapse supernovae”, leaving behind an even more compact final remnant, either a neutron star or black hole, as we discuss next.

20 Post-Main-Sequence Evolution: High-Mass Stars

20.1 Multiple shell burning and horizontal loops in H-R diagram

The post-main-sequence evolution of stars with higher initial mass, $M > 8M_{\odot}$ has some distinct differences from that outlined above for solar and intermediate mass stars. Upon exhaustion of H-fuel at the end of the main sequence, such stars again expand in radius because of the over-luminosity of H-shell burning. But the high luminosity and high surface temperature on the main sequence means that their stellar envelopes remain radiative even as they expand, never reaching the cool temperatures that force a climb up the Hayashi track. Instead, their evolution tends to keep near the constant luminosity set by $L \sim M^3$ scaling for the star's mass, so evolving horizontally to the right on the H-R diagram.

Since stellar radii scale nearly linearly with mass $R \sim M$, the mean stellar density $\rho \sim M/R^3 \sim M^{-2}$ tends to decline with increasing mass. Thus, even after the core contraction that occurs toward the end of nuclear burning, the core density of high-mass stars never becomes high enough to become electron degenerate. Moreover the higher mass means a stronger gravitational confinement that gives higher central temperature and pressure. This now makes it possible to overcome the increasingly strong electrical repulsion of more highly charged, higher elements, allowing nucleosynthesis to proceed up the period table all the way to Iron, which is the most stable nucleus.

However, each such higher level of nucleosynthesis yields proportionally less and less energy. This can be seen from the plot in figure 20.1 of the binding energy per nucleon vs. the number of nucleons in a nucleus. The jump from H to He yields 7.1 MeV, which relative to a nucleon mass of 931 MeV represents a percentage energy release of about 0.7%, as noted above. But from He to C the release is just $7.7 - 7.1 = 0.6$ MeV, representing an energy efficiency of just 0.06%. As the curve flattens out, the fractional energy release become even less, until for elements beyond Iron, further fusion would require the *addition* of energy.

For such massive stars, the final stages of post-main sequence evolution are characterized by an increasingly massive Iron core that can no longer produce any energy by further fusion. But fusion still occurs in a surrounding series of shells, somewhat like an onion skin, with higher elements fusing in the innermost, hottest shells, and outer shells fusing lower elements, extending to an outermost shell of H-burning (see figure 20.2).

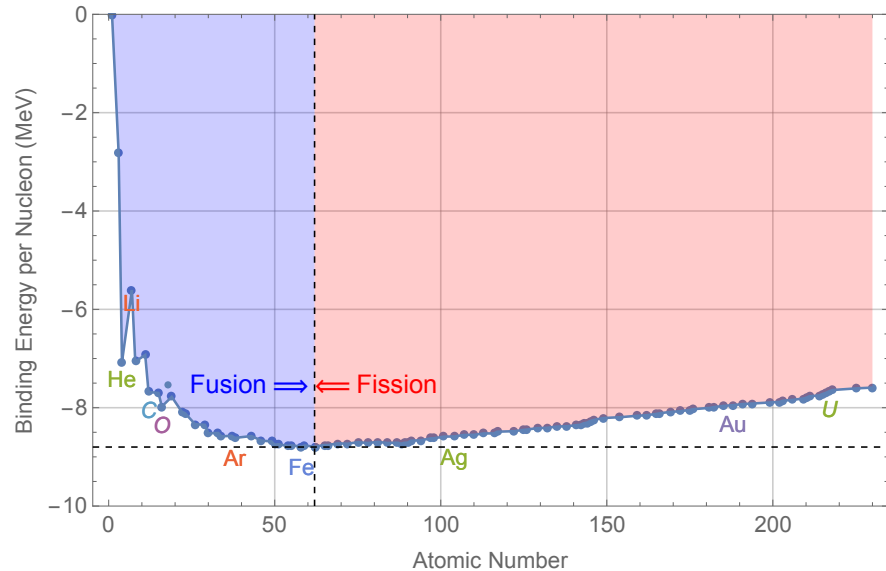


Figure 20.1 Binding energy per nucleon plotted vs. the number of nucleons in a nucleus. The *fusion* of light elements moves nuclei to the right, releasing the energy of nuclear burning in the very hot dense cores of stars, but only up to formation of the most stable nucleus, just beyond Iron (Fe), with atomic number $A = 56$.

Heavier elements are produced in the sudden core collapse of massive-star supernovae, and by merger of binary neutron stars (see figure 20.6). The *fission* of such heavy elements leads to lower-mass nuclei toward the left. The energy released is what powers nuclear fission reactors here on Earth. Adapted from original graphic by Keith Gibbs at <http://www.schoolphysics.co.uk/>.

20.2 Core-collapse supernovae

With the build-up of Iron in the core, there is an increasingly strong gravity, but without the further fusion-generated energy to keep the temperature high, the core pressure becomes unable to support the mass above. This eventually leads to a catastrophic *core collapse*, halted only when the electrons merge with the protons in the Iron nuclei to make the entire core into a collection of *neutrons*, with a density so high that they now actually become *neutron degenerate*. The “stiffness” of this neutron-degenerate core leads to a “rebound” in the collapse, with gravitational release from the core contraction now powering an explosion that blows away the entire outer regions of the star, with the stellar ejecta reaching speeds of about 10% the speed of light! This ejecta contains Iron and other heavy elements, including even those beyond Iron that are fused in less than a second of the explosion by the enormous energy and temperatures. While elements up to Oxygen can also be synthesized in low-mass stars, the heavier

~20 Msun star

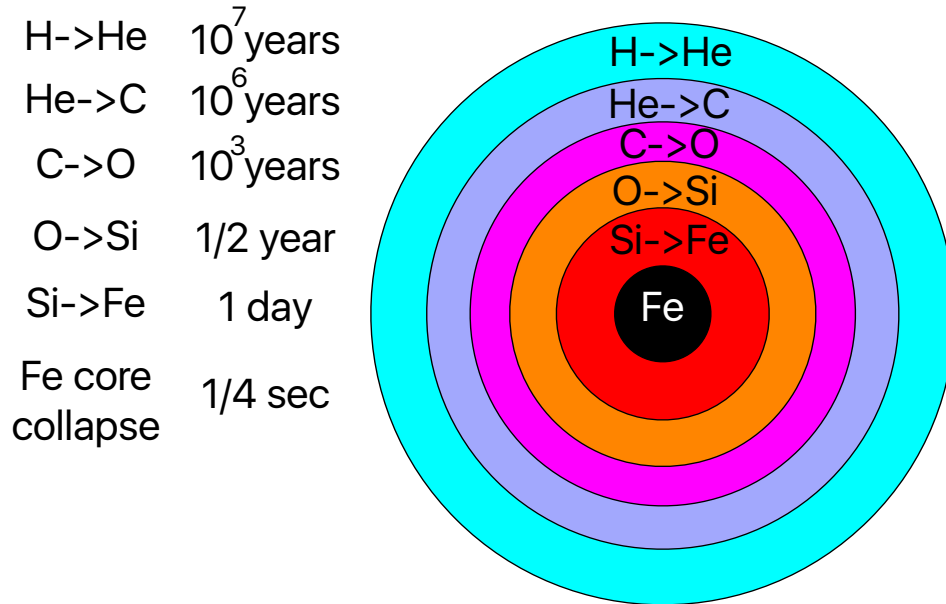


Figure 20.2 The “onion-skin” layering of the core of a $\sim 20M_{\odot}$ star just before supernovae core collapse, illustrating the various stages of nuclear burning in shells around the inert Iron core. The left table shows the decreasing duration for each higher stage of burning.

elements up to Iron are thought to have originated in supernova explosions¹. For a few weeks, the luminosity of such a supernova can equal or exceed that of a whole galaxy, up to $\sim 10^{12}L_{\odot}$!

Though the dividing line is not exact, it is thought that all stars with initial masses $M > 8M_{\odot}$ will end their lives with such a core-collapse supernova, instead of following the track, AGB \rightarrow PN \rightarrow White Dwarf, for stars with initial mass $M < 8M_{\odot}$. Stars with initial masses $8M_{\odot} < M \lesssim 30M_{\odot}$ are thought to leave behind a *neutron star* remnant, as discussed next. But we shall also see that such neutron star remnants have their own upper mass limit of $M_{ns} \lesssim 2.1M_{\odot}$, beyond which the gravity becomes so strong that not even the combination of nuclear forces and degenerate pressure from neutrons can prevent a further collapse, this time forming a *black hole*. This is thought to be the final core remnant for the most massive stars, those with initial mass $M \gtrsim 30M_{\odot}$.

¹ Neutron-rich elements beyond Iron are now believed to be primarily produced in “kilonova” that arise from merger of binary neutron stars.

20.3 Neutron stars

Neutron stars are even more bizarrely extreme than white dwarfs. With a mass typically about twice the Sun's, they have a radius comparable to a small city, $R_{ns} \approx 10$ km, about a factor 600 smaller than even a white dwarf, implying a density that is about 10^8 times higher, and a surface gravity more than 10^5 times higher.

Their support against this very strong gravity comes from both nuclear forces and *neutron* degeneracy pressure, a combination that makes computation of their internal structure very challenging and a topic of much current research. But their overall properties can be well estimated by a procedure for treating neutron degeneracy in way that is quite analogous to that used in §§19.4 and 19.5 for white dwarfs supported by electron degeneracy, just substituting now the electron mass with the *neutron* mass, $m_e \rightarrow m_n \approx m_p$, and setting $Z/A = 1$. The radius-mass relation thus now becomes (cf. eqn. 19.5),

$$R_{ns} = 2^{5/3} \frac{m_e}{m_p} R_{wd} = \frac{1}{GM_{ns}^{1/3}} \frac{\hbar^2}{m_n^{8/3}} \approx \boxed{10 \text{ km} \left(\frac{M_\odot}{M_{ns}} \right)^{1/3}}. \quad (20.1)$$

Note again that, as in the case of an electron-degenerate white dwarf, this neutron-star radius also *decreases* with increasing mass.

For analogous reasons that lead to the upper mass limit for white dwarfs, for sufficiently high mass the neutrons become relativistic, leading now to an upper mass limit for neutron stars (cf. eqn. 19.7) that scales as

$$M_{ns} \leq M_{lim} = 1.1 \left(\frac{\hbar c}{G} \right)^{3/2} \left(\frac{1}{m_p} \right)^2 \approx \boxed{2.1 M_\odot}, \quad (20.2)$$

where again the factor 1.1 comes from detailed calculations not covered here; apart from this and the factor $(A/Z)^2 = 4$, this ‘Tolman-Oppenheimer-Volkoff’ (TOV) limit is the same form as the Chandrasekhar limit for white dwarfs in eqn. 19.7. The exact value of this limiting mass is still a matter of current research, with the value quoted here just somewhat below the value $2.2 M_\odot$ inferred from gravitational waves detected from merger of two neutron stars; see section 20.6. Neutron stars above this limiting mass will again collapse, this time forming a *black hole*.

20.4 Black Holes

Black holes are objects for which the gravity is so strong that not even light itself can escape. A proper treatment requires General Relativity, Einstein's radical theory of gravity that supplants Newton's theory of universal gravitation, and extends it to the limit of very strong gravity. But we can nonetheless use Newton's theory to derive some basic scalings. In particular, for a given mass M , a

characteristic radius for which the Newtonian escape speed is equal to speed of light, $v_{esc} = c$, is just

$$R_{bh} = \frac{2GM}{c^2} \approx 3 \text{ km} \frac{M}{M_\odot}, \quad (20.3)$$

which is commonly known as the “Schwarzschild radius”.

Since the speed of light is the highest speed possible, any object within this Schwarzschild radius of a given mass M can *never escape* the gravitational binding with that mass. In terms of Einstein’s General Theory of Relativity, mass acts to bend space and time, much the way a bowling ball bends the surface of a trampoline. And much as a sufficiently dense, heavy ball could rip a hole in the trampoline, for objects with mass concentrated within a radius R_{bh} , the bending becomes so extreme that it effectively punctures a hole in space-time. Since not even light can ever escape from this hole, it is completely black, absorbing any light or matter that falls in, but never emitting any light from the hole itself. This the origin of the term “black hole”.

Stellar-mass black holes with $M \gtrsim 2.1M_\odot$ form from the deaths of massive stars. If left over from a single star, they are hard or even impossible to detect, since by definition they don’t emit light.

However, in a binary system, the presence of a black hole can be indirectly inferred by observing the orbital motion (visually or spectroscopically via the Doppler effect) of the luminous companion star.

Moreover, when that companion star becomes a giant, it can, if it is close enough, transfer mass onto the black hole. Rather than falling directly into the hole, the conservation of the angular momentum from the stellar orbit requires that the matter first feed an orbiting accretion disk. By the Virial theorem, half the gravitational energy goes into kinetic energy of orbit, but the other half is dissipated to heat the disk, which by the blackbody law then emits it as radiation.

The luminosity of such black-hole accretion disks can be very large. For a black hole of mass M_{bh} accreting mass at a rate \dot{M}_a to a radius R_a that is near the Schwarzschild radius R_{bh} , the luminosity generated is

$$L_{disk} = \frac{GM_{bh}\dot{M}_a}{2R_a} = \frac{R_{bh}}{4R_a} \dot{M}_a c^2 \equiv \epsilon \dot{M}_a c^2. \quad (20.4)$$

The latter two equalities define the efficiency $\epsilon \equiv R_{bh}/4R_a$ for converting the rest-mass-energy of the accreted matter into luminosity. For accretion radii approaching the Schwarzschild radius, $R_a \approx R_{bh}$, this efficiency can be as high as $\epsilon \approx 0.25$, implying 25% of the accreted matter-energy is converted into radiation. By comparison, for H-fusion of a MS star the overall conversion efficiency is about 0.07%, representing the $\sim 10\%$ core mass that is sufficiently hot for H-fusion at a specific efficiency, $\epsilon_H = 0.007 = 0.7\%$.

In the inner disk, the associated blackbody temperature can reach 10^7 K or more (see challenge problem); and at the very inner disk edge, dissipation of the orbital energy can heat material to even more extreme temperatures, up

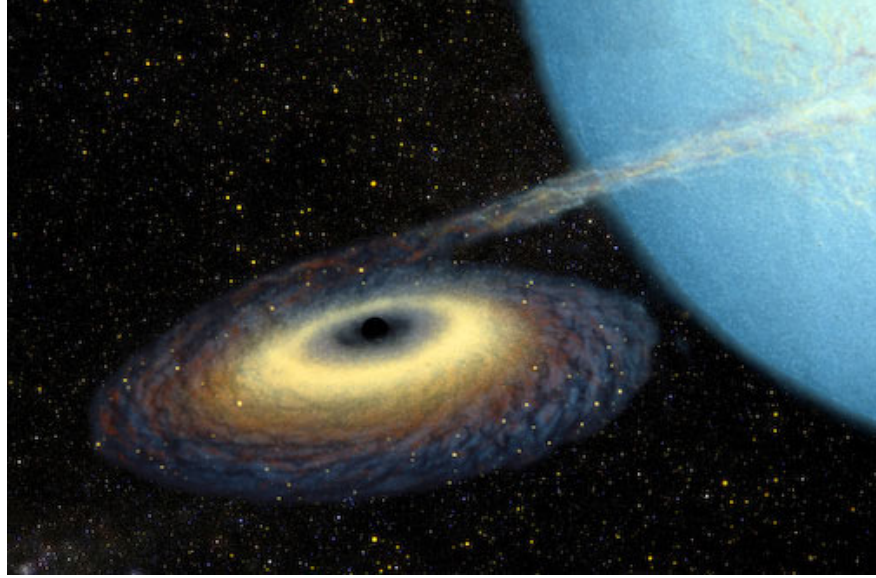


Figure 20.3 Artist depiction of mass transfer onto accretion disk around black hole in Cygnus X-1.

to $\sim 10^{10}$ K. By studying the resulting high-energy radiation, we can infer the presence and basic properties (mass, even rotation rate) of black holes in such binary systems, even though we can't see the black hole itself.

Figure 20.3 shows an artist depiction of the mass transfer accretion in the high-mass X-ray binary Cygnus X-1, thought to be the clearest example of a stellar-remnant black hole, estimated in this case to have a mass $M_{bh} > 10M_{\odot}$ that is well above the $M_{lim} \approx 2.1 M_{\odot}$ upper limit for a neutron star (cf. eqn. 20.2).

20.5 Observations of stellar remnants

It is possible to observe directly all three types of stellar remnants:

1. *Planetary Nebula and White Dwarf stars*

Stars with initial mass $M < 8M_{\odot}$ evolve to an AGB star that ejects the outer stellar envelope to form a Planetary Nebula (PN) with the hot stellar core with mass below the Chandrasekhar mass, $M_{wd} < M_{ch} = 1.4M_{\odot}$. Once the nebula dissipates, this leaves behind a White Dwarf (WD). White dwarf stars are very hot, but with such a small radius that their luminosity is very low, placing them on the lower left of the H-R diagram.

The excitation and ionization of the gas in the surrounding PN makes it shine with an emission line spectrum, with the wavelength-specific emission



Figure 20.4 Left: Optical image of the Crab Nebula, showing the remnant from a core-collapse supernova whose explosion was observed by Chinese astronomers in 1054. Right: A composite zoomed-in image of the central region Crab Nebula, showing the optical (red) image superimposed with an X-ray (blue) image made by NASA's Chandra X-ray observatory. The bright star at the nebular center is the Crab pulsar, a rapidly rotating neutron star that was left over from the supernova explosion. Images courtesy of NASA/Hubble Space Telescope.

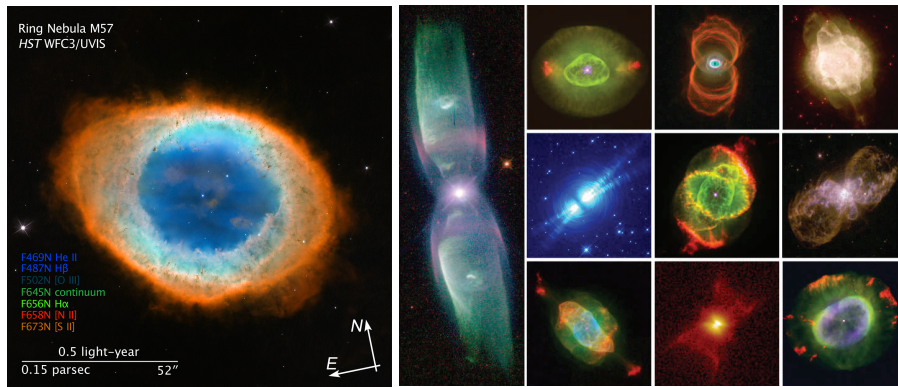


Figure 20.5 Left: M57, known as the Ring Nebula, provides a vivid example of a spherically symmetric Planetary Nebula. The central hot star is the remnant of the stellar core, and after the nebula dissipates, it will be left as a White Dwarf star. The annotations indicate the spectral lines responsible for the various colors, and the lines show the scale and compass orientation of the image. Right: A gallery of planetary nebulae, showing the remarkable variety of shapes that probably stem from interaction of the stellar ejecta with a binary companion, or perhaps even with the original star's planetary system. Images courtesy of NASA/Hubble Space Telescope.

of various ion species giving it range of vivid colors or hues. Figure 20.5 shows that these PN can thus be visually quite striking, with spherical emission

nebula from single stars (left), or very complex geometric forms (right) for stars in binary systems.

2. *Neutron stars and Pulsars*

A star with initial masses in the range $8M_{\odot} < M \lesssim 30M_{\odot}$ ends its life as a core collapse supernova that leaves behind a neutron star with mass $1.4M_{\odot} < M_{ns} < 2.1M_{\odot}$. The conservation of angular momentum during the collapse to such a small size (~ 10 km) makes them rotate very rapidly, often many times a *second*! This also generates a strong magnetic field, and when the polar axis of this field points toward Earth, it emit a strong *pulse* of beamed radiation in the radio to optical to even X-rays. This is observed as a *pulsar*.

One of the best known examples is the Crab pulsar, which lies at the center of the Crab Nebula, the remnant from a core-collapse supernova that was observed by Chinese astronomers in 1054 AD. Figure 20.4 shows images of this Crab nebula in the optical region (left) and in a composite of images (right) in the optical (red) and X-ray (blue) wavebands.

3. *Black holes and X-ray binary systems*

Finally, stars with initial masses $M \gtrsim 30M_{\odot}$ end their lives with a core collapse supernova that now leaves behind a black hole with mass $M_{bh} > 2.1M_{\odot}$. As noted, in single stars, these are difficult or impossible to observe, because they emit no light; but in binary systems, accretion from the other star can power a bright accretion disk around the black hole that radiates in high-energy bands like X-rays and even γ -rays. Figure 20.3 shows an artist depiction of accretion onto a black hole in the high-mass X-ray binary Cygnus X1. As noted in §13.2, the Event Horizon Telescope has recently imaged the mm-wave emission from a supermassive black at the center of the galaxy M87 (see §26.4).

20.6 Gravitational Waves from Merging Black Holes or Neutron Stars

Einstein's publication in 1915 of his General Theory of Relativity cast gravity as the bending of spacetime. It also directly implied that an *accelerating* mass would generate a wave of the changing gravity, propagating at the speed of light. Because gravity is so weak, Einstein himself never thought such waves could be detectable.

Indeed, even for the strongest imagined sources, the merger of two stellar-mass black holes, the expected signal after traveling a characteristic distance to Earth is estimated to be very weak, alternatively stretching and compressing distances by a tiny relative fraction of just 10^{-21} ! Thus, for example, over a length of 1 km, this would require measurement of distance changes on a scale of 10^{-18} m, or about 1/1000 the size of a proton!

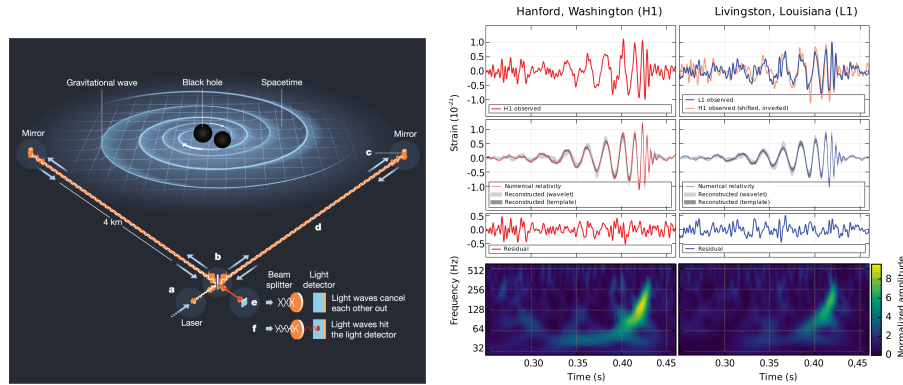


Figure 20.6 *Left:* Illustration for how the merger of orbiting black holes generates a spiral wave in spacetime, which upon propagation through the perpendicular arms of LIGO induces alternate stretches and compressions of the arm lengths that are detected by the interference pattern of two reflected laser beams. *Right:* Actually data traces from the first gravitational wave detection on 17 September 2015, as recorded from stations in Hanford, Washington and Livingston, Louisiana. The lines compare the traces from the two stations with each other, and with numerical models; the bottom color plots show amplitude vs. time and frequency, and the associated frequency increase that gives the characteristic “chirp” when translated into sound. Graphics courtesy Caltech/MIT/LIGO Laboratory.

Nonetheless, through a remarkable combination of ingenuity and heroic scientific ambition, designs were developed over several decades that led to construction in 2002 of an instrument called LIGO (for Laser Interferometer Gravitational-wave Observatory), designed to detect waves from mergers of compact objects like black holes and neutron stars. Following extended further development over more than a decade to improve the sensitivity and reduce noise, a version called Advanced LIGO detected gravitational waves from the merger of two black holes. After traveling over a billion light years, these waves arrived at Earth and the LIGO detectors on 17 September 2015, one century after Einstein’s publication of the General Relativity theory that predicted their existence.

As illustrated figure 20.6, LIGO² was able to detect these tiny deflections by analyzing the interference pattern between two laser signals that reflect from mirrors at the end of two perpendicular, 4 km long arms. Comparing variations from duplicate detectors in both Louisiana and Washington states helped discriminate against false signals from local disturbances and noise sources. As the black holes spiraled ever closer toward merger, the waves steadily increased in frequency, which when translated into sound gave a characteristic “chirp”. (See right panel of figure 20.6.) When analyzed in comparison to computer simulations of such mergers, the chirp pattern indicated the black holes in this first-detected merger were quite massive, about $29M_{\odot}$ and $36M_{\odot}$, several times higher than

² See also the nice video at <https://www.ligo.caltech.edu/video/ligo20160211v1> .

the most-massive black holes inferred in the high-mass X-ray binaries discussed in §20.5. The final, merged black hole was inferred to be about $62M_{\odot}$, with the extra $3M_{\odot}$ converted to energy in the emitted gravitational wave, which for brief, few msec of the merger represented some 50 times the luminosity of all the stars in the observable universe!

Just two years after this historic discovery, the 2017 Nobel prize in Physics was awarded to 3 leaders of the team that developed LIGO. This followed on the 1992 Noble prize, awarded to two astronomers who, in 1982, discovered a binary system that provided strong indirect evidence for gravitational waves. The measured changes in the period of this pulsar orbiting a neutron star closely followed the predicted changes from orbital decay associated with loss of orbital energy through emission of gravitational waves.

In August 2017, LIGO then detected the first *merger* of such *neutron* stars in close binary orbit. Unlike the merger of black holes, which owing to their restriction against any light emission had no detected electromagnetic (EM) signatures, this neutron-star merger was also detected in EM spectral bands ranging from gamma rays and X-rays, to UV and optical light, to infrared and even radio waves. In particular, just 1.7 seconds after the recorded gravitational wave, a 2 second burst of gamma rays was observed by the Fermi and INTEGRAL satellites, which can detect such gamma rays from anywhere in the sky without any directed pointings.

The search for other electromagnetic signals was aided by the localization of the source on the sky, made possible by triangulating the signals of the two LIGO detectors with constraints provided by a third detector called VIRGO in Italy. This led to a plethora of information about both the neutron stars and the associated material ejected from the merger, including spectroscopic signatures of very heavy elements like silver, gold and platinum. This supported earlier suggestions that such high-mass elements, colloquially characterized as “bling” because of their prominent use as precious metals and in jewelry, are mostly produced in the “kilonovae” associated with such neutron-star mergers, rather than, e.g., in core-collapse supernovae of massive stars, as had been previously thought. Figure 20.6 shows a periodic table with our current best estimates for the origin of each element.

Finally, while the energy flux of waves, either in light or gravity, declines with inverse square of distance of the source, the actual wave amplitude only drops with inverse distance. Because LIGO directly detects this wave amplitude, each factor increase in the precision of the detection (from ongoing efforts to reduce the many sources of noise) leads to an equivalent factor increase in the distance that a source of given strength can be detected. But because the volume of space increases with the *cube* of this distance, the number of detectable systems increases with the cube of this improved precision. For example, a factor 2 increase in precision leads to factor 2 in detectable distance, and so a factor 8 in the number of detectable sources. The expectation thus is that, with ongoing

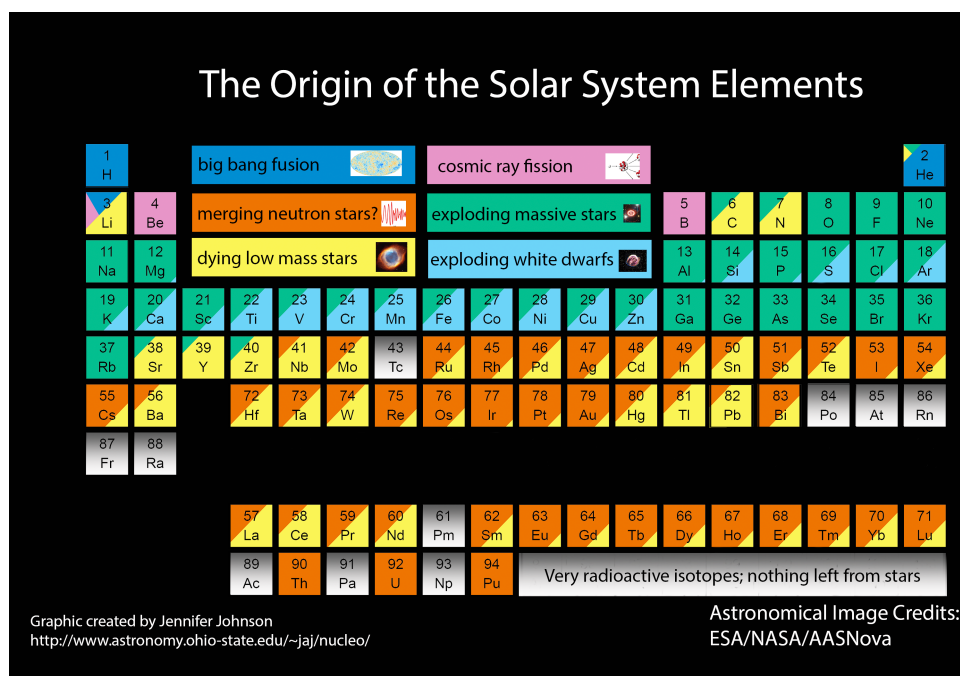


Figure 20.7 Periodic table showing the origin of each of the elements. The orange color shows that most of the heavy, neutron-rich elements, including precious “bling” like gold, platinum, and silver (Au, Pt, Ag), are now understood to have originated from the “kilonova” explosions from merging neutron stars. Graphic courtesy Jennifer Johnson, under Creative Commons Attribution-ShareAlike 4.0 International License. Astronomical image credits to ESA/NASA/AASNova.

and planned improvements to sensitivity and precision, detection rates could approach one new source per day!

20.7 Questions and Exercises

Quick Question 1:

- Because of general relativistic effects, it turns out the lowest stable orbit around a black hole is at radius of $3R_{bh}$. What is the luminosity efficiency for accreting to this radius?
- What is the accretion luminosity, in L_{\odot} , for a mass accretion rate $\dot{M}_a = 10^{-6} M_{\odot}/\text{yr}$ to this radius?
- Challenge problem:** For a black hole with mass $M_{bh} = 3M_{\odot}$, use the Stefan-Boltzmann law to derive the radiative flux that would balance the local gravitational heating at this radius $r = 3R_{bh}$, and then solve for the local blackbody temperature $T(r = 3R_{bh})$. Express this first as a ratio to the Sun’s surface temperature T_{\odot} , and then also in Kelvin.

Exercise 1: Use Wien's law to compute the peak wavelength (in nm) of thermal emission from the inner region of an accretion disk with temperature $T = 10^7$ K. What is the energy (in eV) of a photon with this wavelength? Now also answer both questions for $T = 10^{10}$ K. What parts of the electromagnetic spectrum do these photon wavelengths/energies correspond to?

Part III

Interstellar Medium & Formation of Stars and Planets

21 The Interstellar Medium

21.1 Star-gas cycle

Compared to stars, the region between them, called the *interstellar medium* or “ISM”, is very low density; but it is *not* a completely empty vacuum. For one thing, we’ve seen above that the final remnants of stars, whether white dwarfs, neutron stars, or black holes, generally have much less mass than the initial stellar mass; this implies that a substantial fraction (30-90%) of this initial mass is recycled back into the surrounding ISM through planetary nebulae, stellar winds, or supernova explosions. Moreover, a key theme in this and the next section is that stars are themselves *formed* out of this ISM material through gravitational contraction, making for a kind of star-gas-star cycle, as illustrated in figure 21.1.

If one assumes that, on average, a typical atom spends roughly equal fractions of time in the star vs. ISM phase of this cycle, then the average density of gas in the ISM should be roughly equal to the mass of the stars spread out over the volume between them. For example, in the region of the galaxy near the Sun, the so-called “solar neighborhood”, the mean number density of stars is $n_* \approx 0.1 \text{ pc}^{-3}$, reflecting a typical interstellar separation distance $d \approx n_*^{-1/3} \approx 2 \text{ pc}$. (Recall that the nearest star, α Centauri, is about 1.3 pc from the Sun.) If we take the average mass of each star to be roughly that of the Sun, we obtain a mean mass density $\rho \approx M_\odot n_* \approx 7 \times 10^{-24} \text{ g/cm}^3$.

This is much, much lower than the typical average density within stars, which as noted earlier for the Sun is $\rho_\odot \approx 1.4 \text{ g/cm}^3$; this mostly just reflects the huge distance/size ratio, giving roughly a factor $(\text{pc}/R_\odot)^3 \sim 10^{24}$, for the volume between stars vs. that within them. Thus while most stars typically have mean densities comparable to matter (like water) here on Earth (i.e., $\rho \approx 1 \text{ g/cm}^3$), this ISM density is well below (by a factor $\sim 10^4$) even the most perfect vacuum ever created in terrestrial laboratories ($\rho \sim 10^{-19} \text{ g/cm}^3$).

Indeed, for ISM densities, it is more intuitive and common to quote values in terms of the *atom number density*. For example, with a composition dominated by Hydrogen, the associated ISM Hydrogen-atom number density is $n \approx \rho/m_p \approx 4 \text{ cm}^{-3}$.

The assumptions and approximations behind this estimate – viz. the equal time between ISM and stars, the Sun representing a typical stellar mass, the

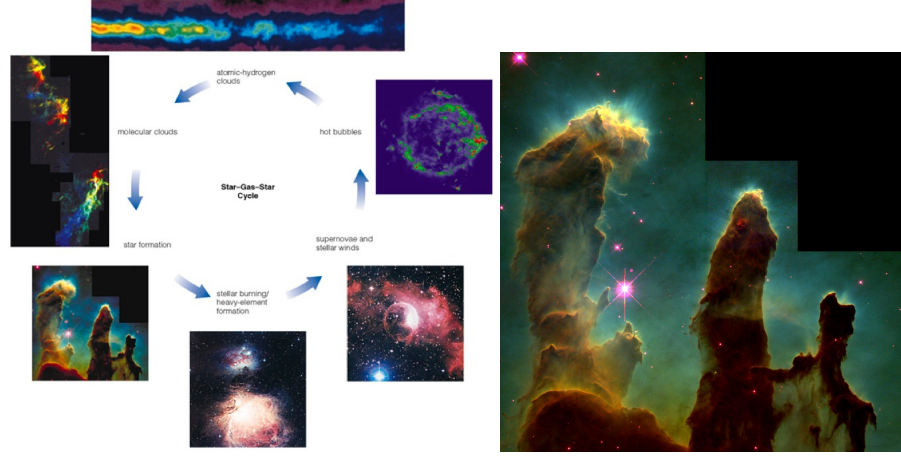


Figure 21.1 *Left:* Illustration of the cycle of mass exchange between stars and the ISM. Starting from top, warm (10^4 K) hydrogen clouds cool to become cold (< 100 K) dense, molecular clouds (left), which undergo gravitational collapse to star formation, as in the “pillars of creation” in the Eagle nebula (lower left). Newly formed massive stars ionize and heat nearby clouds, as in the Orion nebula (bottom), with stellar wind outflows and supernova explosions (lower right) of these short-lived massive stars blowing open hot ISM bubbles (right). Finally, these compress surrounding warm ISM (top), helping induce their cooling to continue the cycle. *Right:* Closer view of “pillars of creation” in the Eagle Nebula, a cold molecular cloud undergoing active star formation and illuminated by recently formed massive stars.

assumed star density – are all somewhat rough, and can even vary through the galaxy. Nonetheless, when averaged over the entire galaxy, the characteristic ISM number density $n \sim 1 \text{ cm}^{-3}$ is indeed comparable to this local estimate. However, within this broad average, there are wide variations, reflecting a highly complex, heterogeneous, and dynamic ISM, as discussed next.

21.2 Cold-Warm-Hot phases of nearly isobaric ISM

A key factor in the wide variations in density of the ISM is the wide variation in its temperature. Roughly speaking, gas in the ISM can be characterized in 3 distinct temperature phases, ranging from cold ($T \sim 10 - 100 \text{ K}$), to warm ($T \sim 5000 - 10,000 \text{ K}$), to hot ($T \sim 10^5 - 10^7 \text{ K}$). Figure 21.2 vividly illustrates the distinct signatures of these different components in various spectral wavebands ranging from the radio to gamma rays, as mapped along the disk plane of our Milky Way galaxy.

In contrast to these wide variations in temperature, the gas *pressure*, which is proportional to the *product* of density and temperature, tends to be relatively constant over the broad ISM, with a typical value $P/k = nT \sim 10^3 \text{ K/cm}^3$. This near constancy of ISM pressure stems from the fact that gravity, which

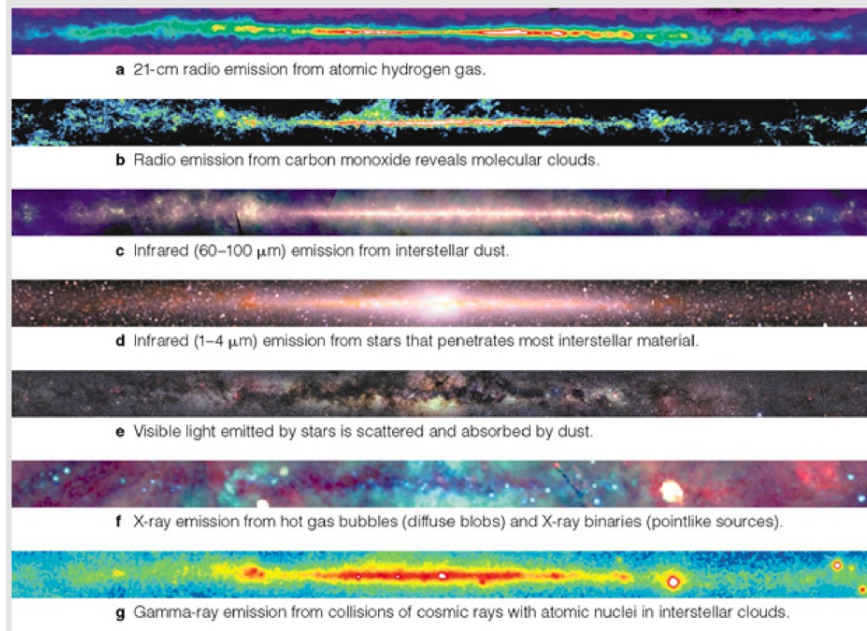


Figure 21.2 Maps along the plane of our Milky Way galaxy, taken in multiple wavebands with energy increasing downward, ranging from 21-cm radio emission (a) at the top, to gamma-rays (g) at the bottom. As annotated in the figure, each waveband is sensitive to distinct components of the multi-phase temperatures of the ISM, along with the disk population of stars in the galaxy. The map is oriented such that the center is toward the galactic center, in the direction of the constellation Sagittarius.

declines in strength with the inverse square of the distance, is generally too weak to confine gas over the many parsec scales between stars¹. In the absence of any restraining force, and ignoring any disturbances from stellar mass ejection (e.g. from stellar winds or supernovae), the ISM gas should over time settle into a dynamical equilibrium that is roughly *isobaric*, meaning with a spatially constant gas pressure.

Within this roughly isobaric ISM, the densities of the 3 phases thus tend to scale inversely with temperature, ranging from $n \approx 10 - 100 \text{ cm}^{-3}$ for cold clouds, to $n \approx 0.1 - 0.2 \text{ cm}^{-3}$ for the warm gas, to $n = 10^{-2} - 10^{-3} \text{ cm}^{-3}$ for the hot component.

Because the flux of radiation also falls with the inverse square of distance, we might expect the temperature of gas far from stars to be always very cold, for example as would be the case for the equilibrium temperature of a blackbody (see QQ 1). But the low density of interstellar gas makes it very different from a

¹ An exception to this is in the densest cloud “cores” in star-forming regions, wherein gravity is compressing cold but very dense and thus high-pressure gas in the final contraction toward forming stars. See § 22.

blackbody, since emitting radiation requires collisions to excite atoms or interact with free electrons, the rates of which decrease with density.

Quick Question 1: a. Recalling that the equilibrium blackbody temperature of the Earth is $T_e \approx (T_\odot/2) \sqrt{2R_\odot/au} \approx 280$ K, show that the corresponding temperature at a distance d from the Sun is given by $T = T_e \sqrt{au/d}$. b. Compute the temperature for $d = 1$ pc. c. Compute the temperature at a distance $d = 1$ pc from a hot star with $T_* = 10 T_\odot \approx 60,000$ K.

Indeed, for the hot ISM the temperature is so high that Hydrogen and Helium are completely ionized, with only the heaviest and most complex atoms, like iron, having a few remaining bound electrons. This means that radiative emission mostly only occurs through rare collisional excitation of these few, partially ionized heavy ions, making radiative cooling very inefficient, with a characteristic cooling time of several Myr. For the warm and still mostly ionized ISM, the higher density and greater number of bound electrons in heavy ions makes cooling somewhat more effective, but cooling times are still quite long, typically of order 10^4 years.

As for the energy source that heats up the hot and warm ISM in the first place, this comes mainly from hot, massive stars. Near such a hot, luminous star, UV photoionization of the surrounding Hydrogen gas can heat it to temperatures that are a significant fraction of the stellar effective temperature, of order 10^4 K. As detailed in §21.4, the resulting volume of ionized gas – dubbed HII regions from the standard notation for ionized hydrogen – can extend several parsecs from the star. Overall, such photo-ionization from hot stars is a significant heating source for the *warm* component of the ISM.

But an even more dramatic source of energy comes from the violent supernova (SN) explosions that end the relatively brief lifetimes of such hot, massive stars. In such SN explosions, several solar masses of stellar material is ejected at very high speeds, approaching 10% the speed of light, implying kinetic energies of order $E_{SN} \sim M_\odot c^2/200 \sim 10^{52}$ erg. As the high-speed, expanding ejecta runs into the surrounding ISM, the resulting shock² wave heats the gas to very high temperatures, initially up to 10^8 K. But as the ISM gas piles up, the expansion slows and cools, ending up with a temperature $T \approx 10^5 - 10^7$ K, with the total pressure comparable to the surrounding ISM. Such SN explosions are thus the primary source of the *hot* component of the ISM.

Reflecting the large expansion of its source SN explosions, this hot ISM component can actually occupy more than half, even up to ~70-80%, of the *volume* of the galaxy. Most of the remaining volume fraction, ~ 15 – 25%, makes up the warm component, with just a relatively small part, $\lesssim 5\%$, being in relatively cool clouds.

And since thermal energy density is just $E_{th} = (3/2)nkT = (3/2)P$, the near

² A shock wave arises whenever two gases collide with supersonic speed. It effectively converts the kinetic energy of the pre-shock gas into heat, yielding post-shock temperatures that scale with specific kinetic energy, or square of the speed, of the pre-shock gas.

constancy of ISM pressure means that the large volume of hot gas also contains most of the ISM thermal energy.

However, in terms of overall distribution of matter, most of the ISM *mass* is in relatively cool, dense clouds. As discussed in § 22, it is these cool clouds that provide the source material for forming new stars, so let us next examine further their nature.

21.3 Molecules and dust in cold ISM: Giant Molecular Clouds

The low temperature and high density of the cold ISM makes it possible for the atoms to combine into molecules, and so much of the cold ISM takes the form of *Giant Molecular Clouds* (GMC). Reflecting the dominant abundance of Hydrogen, the most common molecule is H_2 . Among the heavy elements, carbon monoxide (CO) is usually the most abundant, reflecting its relative stability and the cosmic abundance of both its atomic constituents. Other common molecules include diatomic Oxygen (O_2) and water (H_2O), and in some clouds up to 100 distinct molecular species (including, e.g., alcohol, $\text{CH}_3\text{CH}_2\text{OH}$) have been detected.

The survival, and thus abundance, of GMC molecules requires both a low local gas temperature and low UV flux, both of which become problematic in the vicinity of hot stars. But the high density and low temperature of such GMC also means they tend to have quite high densities of ISM dust, and this can be very effective at shielding the regions from the heating and photo-dissociation by UV light from nearby stars. The dust itself is generally not formed locally, since in even the coldest clouds the density is not high enough for efficient nucleation of microscopic dust grains. Instead it is thought that most dust is formed in the outer layers of cool giant stars, and then blown away into the ISM by a strong stellar wind.

As discussed in § 12, such dust can lead to very strong extinction and reddening of starlight. Figure 21.3 vividly illustrates the heavy extinction of the background starlight by the GMC Barnard 68. Note moreover how the partially extinguished light from stars around the cloud edges is distinctly reddened.

To estimate the dust opacity, note that a spherical dust grain of radius a , mass m_d , and mass density $\rho_d = m_d/(4\pi/3)a^3$ has a physical cross section,

$$\sigma_d \equiv \pi a^2 = \frac{3m_d}{4a\rho_d}. \quad (21.1)$$

The overall opacity of a dust cloud is given by dividing this cross section for an individual dust particle by the mass m_c of cloud material *per* dust particle, i.e. $\kappa_d = \sigma_d/m_c$. For a mass fraction $X_d = m_d/m_c$ of a cloud that is converted into dust, we find using (21.1) that the implied opacity is

$$\kappa_d = \frac{3X_d}{4a\rho_d} \approx 150 \frac{\text{cm}^2}{\text{g}} \frac{X_d}{X_{d,\odot}} \frac{0.1\mu\text{m}}{a} \frac{1\text{g/cm}^3}{\rho_d}. \quad (21.2)$$

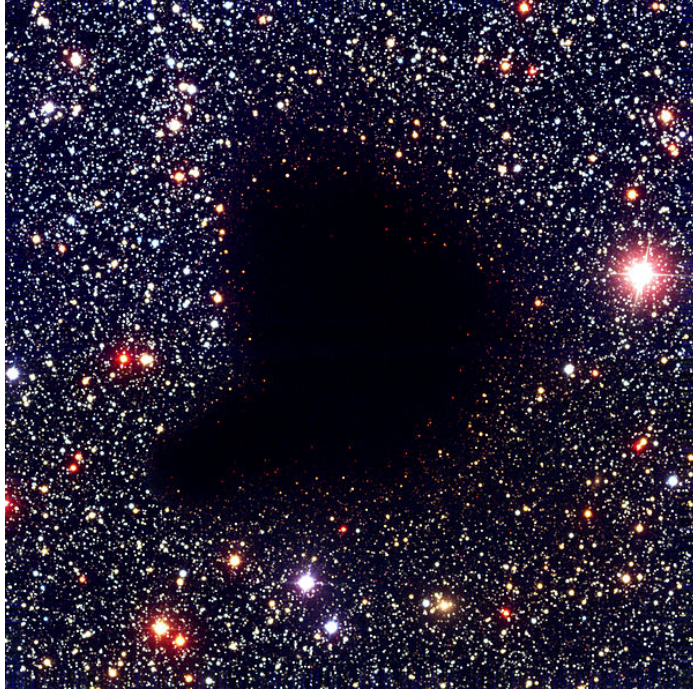


Figure 21.3 Illustration of the heavy extinction of background starlight by the GMC Barnard 68. Note also the reddening of partially extinguished light around the cloud edges. Credit:ESO.

The latter equality provides a numerical evaluation scaled by: the dust mass fraction $X_{d,\odot} \approx 2 \times 10^{-3}$ assuming full conversion of dust-forming material at standard (solar) abundances; the dust grain internal density $\rho_d \sim 1 \text{ g/cm}^3$ (most dust would almost float in water); and a typical dust grain size $a \approx 0.1 \mu\text{m}$. We thus see that the corresponding dust opacity, $\kappa_d \approx 150 \text{ cm}^2/\text{g}$, is several hundred times greater than that for free electron scattering in the fully ionized gas inside a star, $\kappa_e \approx 0.34 \text{ cm}^2/\text{g}$.

As already noted in §12.3, the opacity from the geometric cross section of dust grains only applies to wavelengths comparable to or smaller than the grain size, $\lambda \lesssim a$; for $\lambda > a$, the associated dust opacity decreases as $\kappa_d(\lambda) \sim (\lambda/a)^{-\beta}$, where the power index β is sometimes referred to as the “reddening exponent”. For simple Rayleigh scattering from smooth spheres of fixed size a , $\beta = 4$. In practice, the complex mixtures in sizes and shapes of dust typically lead to a smaller effective power index, $\beta \approx 1 - 2$. Still, the overall inverse dependence on wavelength means that clouds that are optically thick to dust absorption and scattering will show a substantially *reddened* spectrum.

This reddening can be quantified in terms of a formal “*color excess*”, which then can be used to estimate an associated visual extinction magnitude $A_V \equiv V_{\text{obs}} - V_{\text{intrinsic}}$. Recall from §12.3 that the extinction magnitude in any waveband

is related to the associated optical depth, which in turns scales linearly with opacity in that waveband, $\kappa(\lambda)$. For wavelengths larger than the dust size $\lambda > a$, and assuming a linear reddening exponent $\beta \approx 1$, we thus see that the extinction magnitude declines with the inverse of the wavelength,

$$A(\lambda) \sim \tau(\lambda) \sim \kappa(\lambda) \sim 1/\lambda. \quad (21.3)$$

For example, if a star has an extinction A_V in the visual waveband centered on $\lambda_V \approx 500 \text{ nm}$, then in the mid-infrared “M-waveband” at roughly a factor ten higher wavelength $\lambda_M \approx 5000 \text{ nm} = 5 \mu\text{m}$, the opacity, and thus the optical depth and extinction magnitude, are all *reduced* by this same factor 10, $A_M \approx A_V/10$. For a case with, say $A_V = 10.8$ magnitudes of visual extinction, the visual flux would be reduced by a factor $e^{-\tau_V} = e^{-A_V/1.08} = e^{-10} = 4.5 \times 10^{-5}$. By contrast, in this mid-IR M-band, the factor ten lower extinction magnitude $A_M = 1.08$ implies a much weaker reduction, now just a factor $e^{-\tau_M} = e^{-A_M/1.08} = e^{-1} = 0.36$.

Stars are typically formed out of interstellar gas and dust in very dense molecular clouds, which often have 10 or 20 magnitudes of visual extinction ($A_V \approx 10 - 20$), essentially completely obscuring them at visual wavelengths. But such stars can nonetheless be readily observed with minimal extinction in mid-IR (few microns) or far-IR (millimeter) wavebands. As discussed in §13, this fact has spurred efforts to build large infra-red telescopes, both on the ground and in space. The ground-based telescopes are placed at high altitudes of very dry deserts, to minimize the effect of IR absorption by water vapor in the Earth’s atmosphere. Another issue is to keep the IR detectors very cold, to reduce the thermal emission background.

Finally, the energy from dust-absorbed optical or UV light is generally reemitted in the mid-IR, at wavelengths set by the dust temperature through roughly the standard Wien’s law for peak emission of a blackbody, $\lambda_{max} \approx 30 \mu\text{m}/(T/100 \text{ K})$ (cf. eqn. 4.6). For GMC clouds with $T = 30 - 50 \text{ K}$ this gives thermal dust emission in the 60-100 μm range, as illustrated for galactic plane dust emission in figure 21.2c.

Quick Question 2: *GMC dust extinction and reddening*

- For a GMC with molecular Hydrogen density $n = 100 \text{ cm}^{-3}$, compute the associated mass density ρ .
- For UV light with $\lambda = 100 \text{ nm}$ and dust with size $a = 0.1 \mu\text{m}$ and solid density $\rho_d = 1 \text{ g/cm}^3$ and the solar abundance mass fraction $X_d = 2 \times 10^{-3}$, use the geometric cross section opacity derived in the text to compute the mean free path ℓ (in pc) for this GMC.
- For a GMC of diameter $D = 30 \text{ pc}$, compute the optical depth τ and reduction fraction F_{obs}/F for a star behind the cloud that emits such UV light.
- Use this to compute the associated extinction magnetic for this UV light, A_{UV} .
- Assuming a reddening exponent $\beta = 1$, now compute the extinction A_V for visible light with $\lambda = 500 \text{ nm}$, and the extinction A_{NIR} for near IR light with $\lambda = 2 \mu\text{m}$.

21.4 HII regions

Let us next consider the warm ISM that is heated by UV photo-ionization from hot, luminous OB-type stars. Specifically, consider an ISM cloud with a uniform number density n of Hydrogen atoms surrounding a hot star with luminosity L . Stellar UV photons with energy $h\nu > h\nu_o \equiv 13.6 \text{ eV}$ can efficiently ionize neutral Hydrogen atoms, but these will then tend to quickly *recombine* with the free electrons.

The recombination rate scales with the electron density n_e times a temperature-dependent recombination coefficient,

$$\frac{1}{t_r} = n_e \langle \sigma_r v_e \rangle_T \approx 4 \times 10^{-13} (\text{cm}^3/\text{s}) n_e \quad ; \quad \text{for } T \approx 10^4 \text{ K}, \quad (21.4)$$

where t_r is the recombination time for an ionized Hydrogen atom, i.e. the time for a free proton to encounter and recombine with a free electron. The recombination coefficient depends on the recombination cross section σ_r times the electron thermal speed v_e , with the angle brackets representing averaging over the thermal velocity distribution of electrons at a temperature T . As mentioned above, such UV photoionization tends to heat the gas to a temperature $T \approx 10^4 \text{ K}$, and so the latter relation evaluates this recombination coefficient for that temperature.

The total emission rate of stellar UV ionizing photons can be estimated by integrating over the Planck blackbody function B_ν from the ionization threshold frequency ν_o , where $h\nu_o \equiv 13.6 \text{ eV}$,

$$\dot{N}_{UV} \equiv L \int_{\nu_o}^{\infty} \frac{B_\nu d\nu}{B h \nu} = \frac{L}{B h} \int_{\nu_o}^{\infty} \frac{B_\nu d\nu}{\nu}. \quad (21.5)$$

Here B is the spectrally integrated Planck function, and the division by the photon energy $h\nu$ converts the energy rate into a photon *number* rate.

In equilibrium, this number of ionizing photons will balance the total number of recombinations over a sphere (commonly dubbed a Strömgren sphere after the scientist who first described it) of radius R_S centered on the star. In terms of the proton (i.e., ionized H) number density n_p , each of the total number $n_p (4/3)\pi R_S^3$ of ionized H in the sphere recombines with an electron of number density n_e over the recombination time t_r . The balance with stellar ionizing photons of emission rate \dot{N}_{UV} thus requires,

$$\dot{N}_{UV} = \frac{4\pi n_p R_S^3}{3t_r} = n_p n_e \langle \sigma_r v_e \rangle_T \frac{4\pi}{3} R_S^3. \quad (21.6)$$

For full ionization of a pure H cloud of number density n , we have $n_p = n_e = n$; thus the “Strömgren radius” R_S of such an “HII region” of ionized Hydrogen (HII) is simply given by

$$R_S = \left[\frac{3\dot{N}_{UV}}{4\pi n^2 \langle \sigma_r v_e \rangle_T} \right]^{1/3} \approx 6.0 \text{ pc} \left[\frac{\dot{N}_{50}}{n_2^2} \right]^{1/3}, \quad (21.7)$$

where $\dot{N}_{50} \equiv \dot{N}_{UV}/(10^{50} \text{ s}^{-1})$ and $n_2 \equiv n/(10^2 \text{ cm}^{-3})$ are convenient variables scaled by typical values for this photon rate and Hydrogen number density.

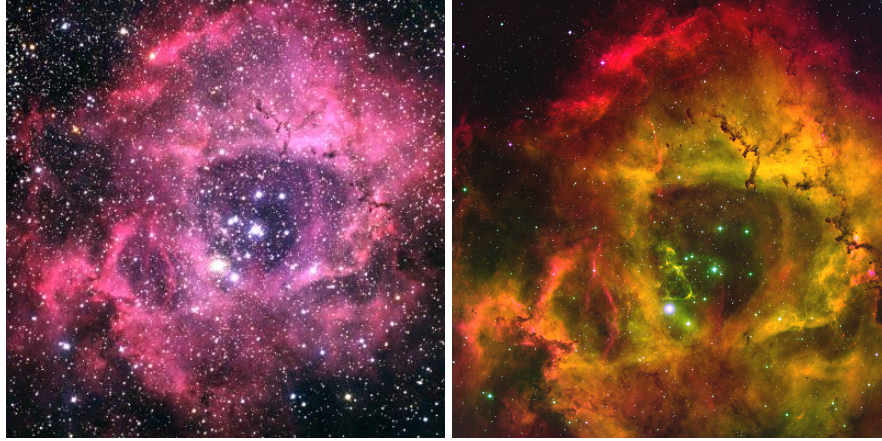


Figure 21.4 *Left:* True-color optical image of the Rosetta nebula and its associated HII region. The reddish glow is from $H\alpha$ line emission from recombination of the ionized Hydrogen. The central cavity has been evacuated by the strong, high-speed stellar winds from the central hot star. *Right:* Composite false-color image showing the emission in $H\alpha$ (red), and lines of OIII (green) and SII (blue). Credit: NASA/HST

The number of UV photons can be estimated from the spectral type of the exciting star, and the number density of H atoms can be inferred from the observed line emission from the HII region. Using these to compute the physical size R_S , we can then use the measured angular radius α to estimate the HII region's distance, $d = R_S/\alpha$.

In an HII region the ongoing recombination occurs through a cascade of electrons from higher to lower bound states of Hydrogen, leading to extensive emission lines for all the Hydrogen term series (Lyman, Balmer, Paschen etc. for lower final state $n=1, 2, 3$, etc.). But in optical images, the most prominent line emission stems from the $n=3$ to $n=2$ transition associated with the Balmer line $H\alpha$, which is in the *red* part of the visible spectrum, with wavelength $\lambda = 656.28 \text{ nm}$. Viewed in the visible, HII regions thus generally have a distinctly reddish glow, as illustrated in the left panel of figure 21.4 for the HII region known as the Rosetta nebula. But the false-color image of the same nebula in the right panel shows that, in addition to the $H\alpha$ (now color-coded red), there is also line emission from doubly ionized Oxygen (OIII, green) and singly ionized Sulfur (SII, blue).



Figure 21.5 *Left:* Hubble space telescope optical image of M51, the “Whirlpool galaxy”. The reddish blotches are from Balmer series $H\alpha$ line emission (which at wavelength $\lambda = 656 \text{ nm}$ is in the red part of the visible spectrum) from giant HII regions. These represent the merger of many individual HII regions that arise when dense regions of interstellar Hydrogen in otherwise cold giant molecular clouds (GMC) are photo-ionized by the UV radiation from the numerous, recently formed, hot massive stars. Note their proximity to dark bands formed from absorption of background stellar light by cold interstellar dust, which outline the galactic spiral arms. *Right:* Composite image of M51 from 4 NASA orbiting telescopes. X-rays (purple) detected by the Chandra X-ray Observatory reveal point-like sources from black holes and neutron stars in binary star systems, as well as a diffuse glow of hot ISM gas. Optical data from the Hubble Space Telescope (green) and infrared emission from the Spitzer Space Telescope (red) both highlight long lanes in the spiral arms that consist of stars and gas laced with dust. Finally, UV light (blue) from the GALEX telescope comes from hot, young stars, showing again how well these track the HII giants and star-forming GMCs along the spiral arms.

21.5 Galactic organization of ISM and star-gas interaction along spiral arms

In the dense regions of active star formation in the Milky Way and other galaxies, the ionization from numerous young, hot, massive stars can merge into an extended *Giant HII region*. Viewed from Earth along the plane of the Milky Way, the projection of foreground and background stars and nebulae can make such regions appear complex and amorphous. A visually clearer view can be gleaned from external galaxies that are viewed *face on*, like the “Whirlpool” galaxy (M51) shown in figure 21.5. The distinctly reddish splotches seen in the optical image in the left panel are all Giant HII regions that formed in the dense clouds along this galaxy’s spiral arms.

The right panel of figure 21.5 shows a composite image in 4 distinct spectral bands, spanning the IR (red), optical (green), UV (blue), and X-rays (purple). The face-on view nicely complements the disk-embedded perspective images from multiple wavebands shown for our Milky Way galaxy in figure 21.2. Note in particular how the close link between ISM and star formation is organized by

the spiral arm structure. This is discussed further in the section on external galaxies (§ 27).

22 Star Formation

22.1 Jeans Criterion for gravitational contraction

Stars generally form in clusters from the gravitational contraction of a dense, cold GMC. The requirements for such gravitational contraction depend on the relative magnitudes of the total internal thermal (kinetic) energy K versus the gravitational binding energy U . For a cloud of mass M , uniform temperature T , and mean mass per particle μ , the total number of particles $N = M/\mu$ have an associated total thermal energy,

$$K = \frac{3}{2} N k T = \frac{3}{2} \frac{M k T}{\mu}. \quad (22.1)$$

If the cloud is spherical with radius R and uniform density $\rho = \mu n = M/(4\pi R^3/3)$, the associated gravitational binding energy (cf. eqn. 8.3) is

$$U = -\frac{3}{5} \frac{G M^2}{R}. \quad (22.2)$$

Recalling the condition $K = -U/2$ for stably bound systems in *virial* equilibrium, we can expect that for a cloud with $K > -U/2$, the excess internal pressure would do work to expand the cloud against gravity, leading it to be less tightly bound (or even unbound, if $K > -U$).

Conversely, for $K < -U/2$, the too-low pressure would allow the cloud to gravitationally contract, leading to a more strongly bound cloud. The critical requirement, known as the *Jeans criterion*, for such gravitational contraction can thus be written

$$\boxed{\frac{M}{R} > \frac{5kT}{G\mu}}. \quad (22.3)$$

In terms of the cloud's atomic number density $n = \rho/\mu = N/(4\pi R^3/3)$, we can define a minimal *Jeans radius* for cloud contraction,

$$\boxed{R_J} \approx \left(\frac{15kT}{4\pi n G \mu^2} \right)^{1/2} \approx 9.6 \text{ pc} \left(\frac{T}{n} \right)^{1/2} \frac{m_p}{\mu}, \quad (22.4)$$

where the second equality assumes CGS units, with number density n in cm^{-3} and temperature T in Kelvin.

Alternatively, one can define a minimum *Jeans mass* (the total mass within a Jeans radius) for a cloud to contract,

$$M_J \equiv \frac{4\pi R_J^3}{3} \mu n \approx \frac{5}{\mu^2} \left(\frac{kT}{G} \right)^{3/2} \left(\frac{15}{4\pi n} \right)^{1/2} \approx 92 M_\odot \frac{T^{3/2}}{n^{1/2}} \left(\frac{m_p}{\mu} \right)^2. \quad (22.5)$$

For typical ISM conditions, both the Jeans radius and mass are quite large, implying it can be actually quite difficult to initiate gravitational contraction. For example, for a cold cloud with $\mu = m_p$, $T = 100$ K and $n = 10 \text{ cm}^{-3}$, we find $R_J \approx 30$ pc and $M_J \approx 30,000 M_\odot$, requiring then a cloud that is initially extremely large and massive. The requirements are somewhat less severe once the hydrogen atoms form into H_2 molecules, thus increasing the molecular weight to $\mu \approx 2m_p$, and so reducing R_J by a factor 2, and M_J by a factor 4.

But a general upshot of such a large Jeans mass is that stars tend typically to be formed in large clusters, resulting from an initial contraction of a GMC, with mass of order $10^4 M_\odot$ or more.

Exercise 1:

- Assuming an isobaric ISM with the canonical pressure $P/k = nT = 10^3 \text{ K cm}^{-3}$, derive expressions for R_J (in pc) and M_J (in M_\odot) as a function of temperature T (in K).
- Now derive analogous expressions for R_J and M_J as a functions of number density n (in cm^{-3}).

22.2 Cooling by molecular emission

In contrast to the poor radiative efficiency of the ionized gas in the warm and hot phases of the ISM, in the cool ISM the formation of molecules makes such clouds much more efficient for radiative cooling. The thermal, collisional excitation of the molecules and dust leads to emission of radiation at IR wavelengths comparable to those associated with black-body emission for the given temperature. For example, for a cloud with temperature $T = 100$ K, radiation is at IR wavelengths $\lambda \approx \lambda_{\text{max},\odot} T_\odot / T \approx 30 \mu\text{m}$ (see eq. 4.6).

At low temperatures $T < 100$ K cooling by molecular radiation is dominated by *carbon monoxide* (CO). Both C and O are relatively abundant elements, and the molecular structure of CO provides a variety of excitation modes (rotational, vibrational, or electronic) from inelastic collision with molecular hydrogen. This converts kinetic energy of the gas to potential energy in the molecules, which de-excite radiatively to emit an IR photon that escapes the cloud, causing it to cool.

Such CO molecular cooling is a key factor in initiating and maintaining cloud contraction, by allowing the cloud to shed the increased internal energy gained from the tighter gravitational binding. In virial equilibrium only half this energy is lost, and so the interior would still heat up in proportion to the stronger gravitational binding. But in practice CO emission is often so efficient that the

cloud interior can stay cool, or even become cooler, as it contracts. The resulting dramatic reduction in interior pressure support then leads to a full *gravitational collapse*.

22.3 Free-fall timescale and the galactic star formation rate

To estimate the timescale for gravitational collapse, recall first from Kepler's third law (see eqn. 10.5) that the period P for orbit at radius R around an object of mass M is

$$P = \sqrt{\frac{4\pi^2 R^3}{GM}} = \sqrt{\frac{3\pi}{G\rho}}, \quad (22.6)$$

where the second equality casts this in terms of the mean density within a sphere of this radius, $\rho = M/(4\pi R^3/3)$. Since the period from Kepler's law does not depend on the orbital eccentricity, (22.6) also applies to a purely radial orbit (with eccentricity $\epsilon = 1$) through a central point mass from this radius. But the self-gravitational collapse of a cloud would occur at just this same rate, implying then that the *free-fall time* to contract to zero radius from the initial radius R should be just a quarter of this orbital period,

$$t_{\text{ff}} = \frac{P}{4} = \sqrt{\frac{3\pi}{16G\rho}} = \frac{0.82 \text{ hr}}{\sqrt{\rho}} = \frac{51 \text{ Myr}}{\sqrt{n}} \sqrt{\frac{2m_p}{\mu}}, \quad (22.7)$$

where the density evaluations assume CGS units (i.e., g/cm^3 for ρ and cm^{-3} for n). For a star like the Sun, for which the mass density is CGS order unity, free fall would be less than an hour. But for a cold molecular cloud, with say a number density $n \approx 100 \text{ cm}^{-3}$, such free-fall would take several Myr.

Quick Question 1: *Free-fall time for simple constant-gravity model*

Recall the elementary physics result that an object falling under gravitational acceleration g drops a distance $s = gt^2/2$ in time t . Fixing the gravity at a constant value $g = GM/R^2$, use this simple relation to solve for the time $t_g(R)$ to fall through a stellar radius (i.e., by setting $s = R$). Compare this with the free fall time in (22.7) by evaluating the ratio $t_g(R)/t_{\text{ff}}$.

In our galaxy, the total mass in giant molecular clouds with density $n \gtrsim 100 \text{ cm}^{-3}$ is estimated to be about $M_{\text{GMC}} \approx 10^9 M_{\odot}$. Since this mass should collapse to stars over a free-fall time, it suggests an overall galactic star formation rate should be given by

$$\dot{M}_{\text{sfr}} = \frac{M_{\text{GMC}}}{t_{\text{ff}}} \approx 200 \frac{M_{\odot}}{\text{yr}}. \quad (22.8)$$

But the observationally inferred star formation rate is actually much smaller, only about $1 M_{\odot}/\text{yr}$, implying an effective *efficiency* of only $\epsilon_{\text{ff}} \lesssim 0.01$. The reasons for this are not entirely clear, but may stem in part from inhibition of gravitational collapse by interstellar magnetic fields, and/or by interstellar

turbulence. Another likely factor is the *feedback* from hot, massive stars, which heat up and ionize the cloud out of which they form, thus inhibiting the further gravitational contraction of the cloud into more stars.

22.4 Fragmentation into cold cores and the Initial Mass Function (IMF)

In those portions of a GMC that do undergo gravitational collapse, the contraction soon leads to higher densities, and thus to smaller Jeans mass and Jeans radius, along with a shorter free-fall time. This tends to cause the overall cloud, with total mass $10^4 - 10^6 M_\odot$, to *fragment* into much smaller, stellar-mass cloud “cores” that will form into individual stars.

A key, still-unsolved issue in star formation regards the physical processes and conditions that determine the *mass distribution* of these proto-stellar cores, leading then to what’s known as the stellar *Initial Mass Function* (IMF).

This IMF can be written as dN/dm , wherein $m = M/M_\odot$ is the stellar mass in solar units, and $dN(m) = (dN/dm)dm$ represents the fractional number of stars within a mass range m to $m + dm$. Studies of the evolution of stellar clusters suggest that this can be roughly characterized by a power-law form,

$$\boxed{\frac{dN}{dm} = Km^{-\alpha}}, \quad (22.9)$$

where K is a normalization factor that depends on the total number of stars, and the power index α has distinct values for high-mass vs. low-mass stars. For $m > 1$, the most commonly inferred value is $\alpha \approx 2.35$, known as the “Salpeter” IMF (dashed red line in figure 22.1), after the scientist who first quantified the concept of an IMF. The large power-index reflects the fact that higher-mass stars are much rarer than lower-mass stars.

For lower mass, there are various models, the simplest being the “flattened” IMF (from Scalo 1986, dashed blue curve in figure 22.1), with $\alpha = 0$ for $m < 1$. Another is the three-power model (from Kroupa 2001, aqua blue curved in figure 22.1), which keeps $\alpha = 2.35$ for $m > 0.5$, but then takes $\alpha = 1.3$ for Red dwarf stars in the mass range $0.08 < m < 0.5$, and $\alpha = 0.3$ for Brown dwarf stars with $m < 0.08$. Figure 22.1 compares various other IMFs.

Exercise 2: Flattened Salpeter IMF

- For the simple flattened Salpeter IMF, with $\alpha = 2.35$ for $m > 1$ and $\alpha = 0$ for $m < 1$, integrate (22.9) over all masses to obtain an expression for the normalization K in terms of the total number of stars N_{tot} .
- Now use this to obtain an expression for the *fraction* of stars, $N(m > m_o)/N_{tot}$, with mass greater than some mass lower limit m_o (assuming $m_o \geq 1$). In particular, what fraction of stars have $m > 1$?

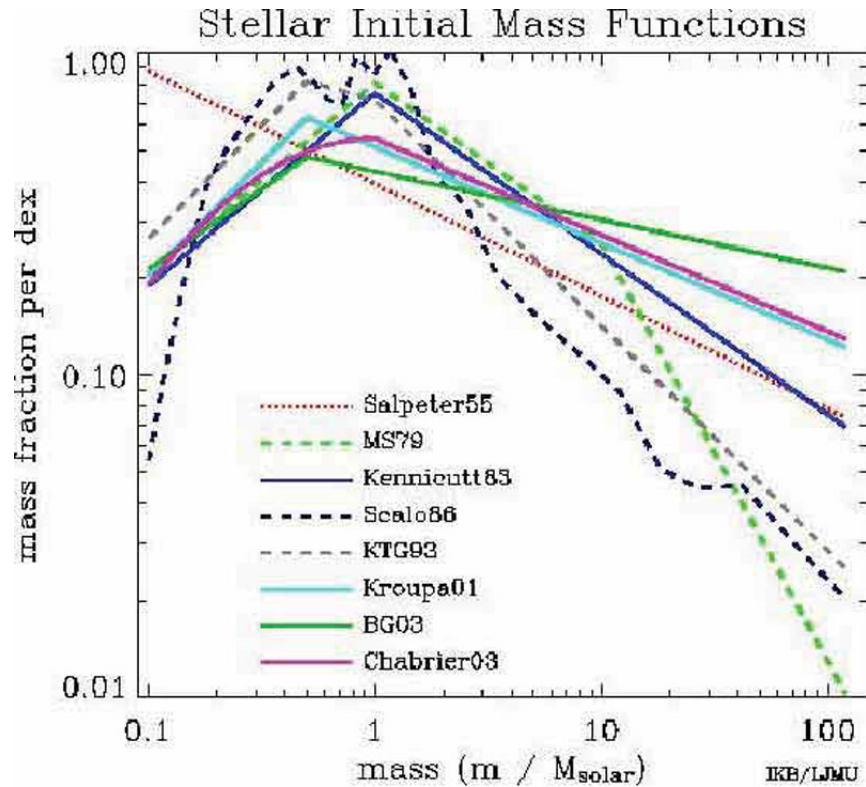


Figure 22.1 Comparison of various IMFs, plotted on a log-log scale in terms of the mass fraction per logarithm interval of mass (a.k.a. “dex”), which is proportional to $m^2 dN/dm$.

c. For $m_o = 100$, how many total stars must a cluster have for there to be at least one star with $m \geq m_o$? How about for $m_o = 300$? What does this imply for observational efforts to determine whether there is an upper mass cutoff to the IMF?

With a given form of the IMF for a collapsing GMC, one can model the evolution of the resulting stellar cluster, based on how each star with a given mass evolves through its various evolutionary phases, e.g. main sequence, red giant, etc.

22.5 Angular momentum conservation of rotating cores and disk formation

In general, the fragmentation of a GMC into stellar-mass cores will endow those cores with a non-zero rotation, and this can be a key factor in their final collapse toward stellar size. While material near and along the core rotation axis can still collapse to form the central star, the conservation of angular momentum

for material near the rotational equator can halt the contraction and lead to formation of a *protostellar disk*.

For material with angular-momentum-per-unit-mass $j \equiv vr$ in circular orbit with speed v at a radius r about a central mass M , the orbital condition for balance between centrifugal and gravitational acceleration can be cast in the form,

$$\frac{GM}{r^2} = \frac{v^2}{r} = \frac{j^2}{r^3} \quad (22.10)$$

For an initially spherical core with starting radius R and angular rotation frequency Ω , a mass parcel at the rotational *equator* has an angular momentum per unit mass $j_{eq} = \Omega R^2$. As the cloud collapses under the gravitational attraction of its own mass M , conservation of angular momentum causes this parcel to rotate faster until it reaches the condition (22.10) for orbit, with an associated “disk” radius

$$r_d = \frac{j_{eq}^2}{GM} = \frac{\Omega^2 R^4}{GM} \equiv 2\beta_{eq}R. \quad (22.11)$$

The last equality here introduces the initial equatorial ratio of rotational to gravitational energy,

$$\beta_{eq} \equiv \frac{\Omega^2 R^2/2}{GM/R} = \frac{3\Omega^2}{8\pi G\rho} = \frac{\Omega^2 P_{orb}^2}{8\pi^2} = \frac{1}{2} \frac{\Omega^2}{\Omega_{orb}^2}. \quad (22.12)$$

The second equality here shows that β_{eq} depends only on the core density ρ and its rotation frequency Ω , two quantities that can generally be readily inferred from observations, with observed cloud cores typically giving $\beta_{eq} \approx 0.02$. For a typical observed core size $R \approx 0.05$ pc, the expected disk radius r_d is a few hundred au, comparable to the inferred sizes of protostellar disks.

The last two equalities recast β_{eq} in terms of the orbital period P_{orb} or orbital frequency $\Omega_{orb} \equiv 2\pi/P_{orb}$.

Initially such disks can have a mass that is a substantial fraction of that for the central star. But in disks with Keplerian orbits, the orbital frequency increases inward with radius as $\Omega_{orb} \sim r^{-3/2}$, meaning that between two neighboring rings there is an overall *shear* in orbital speed. Any frictional interaction – e.g., due to *viscosity* – between such neighboring rings will thus tend to transport angular momentum from the faster inner ring to the slower outer ring, allowing the inner mass to fall further inward, while the angular momentum receiving material moves further outward. Since the specific angular momentum increases outward as $j = vr = \Omega r^2 \sim \sqrt{r}$, this outward viscous diffusion of angular momentum allows over time for most of the mass to accrete onto the star, with just a small mass fraction retaining the original angular momentum. Eventually this remnant disk-mass can fragment into its own gravitationally collapsing cores to form planets. In our own solar system the most massive planet Jupiter has only 0.1% the mass of the Sun, but 99% of the solar system’s angular momentum.

Of course Earth too originated from the evolving proto-solar disk. You and I

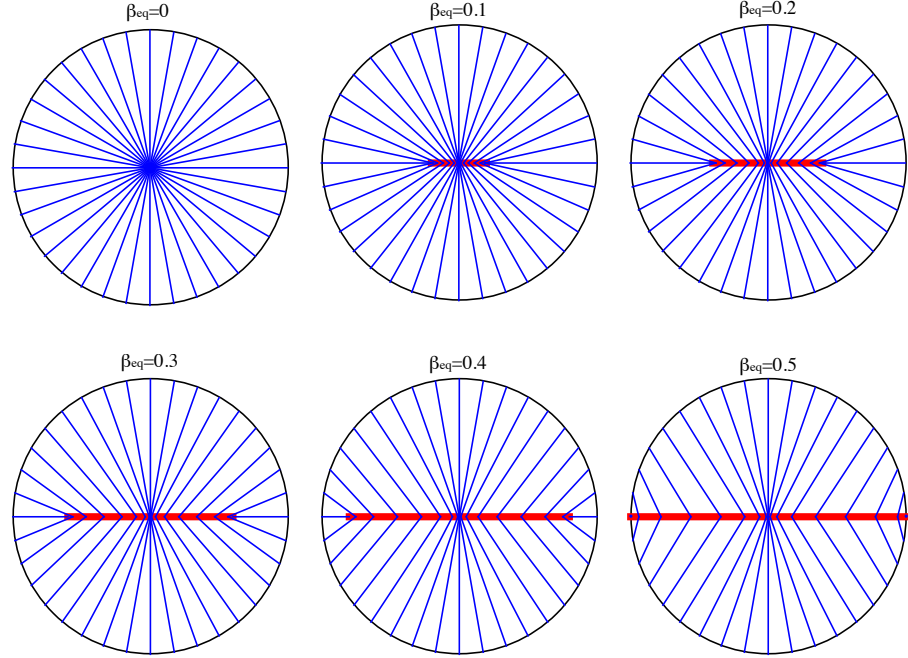


Figure 22.2 Traces (blue lines) illustrating how conservation of angular momentum causes various locations on the surface of a rigidly rotating spherical cloud (represented by black circle) to collapse onto an orbiting disk (marked in red). The various panels are for the labeled values β_{eq} of the equatorial rotational energy to gravitational energy. Note how material near the rotational poles contracts to the concentrated central region, while material at lower latitudes near the equator collapses onto the orbiting disk with outer radius $r_d = 2\beta_{eq}R$, as given by (22.11).

and everyone on Earth are here today because our source material happened to stem from the equatorial regions of the proto-solar core, with too much angular momentum to fall into the Sun itself; it also then could have been the viscous recipient of the angular momentum from other proto-solar-disk material that did diffuse inward onto the Sun. The formation of such planetary systems around our Sun and other stars is discussed in the next section (§23).

22.6 Questions and Exercises

Quick Question: What is the free-fall time for a star like the Sun?

Quick Question: What is the free-fall time for a GMC with number density $n = 100 \text{ cm}^{-3}$ of molecular Hydrogen?

Exercise 3: *Disk collapse from various latitudes*

a. For a spherical cloud of radius R and rotation frequency Ω , consider locations away from the equator, with co-latitude θ measured from the polar axis. Derive an expression

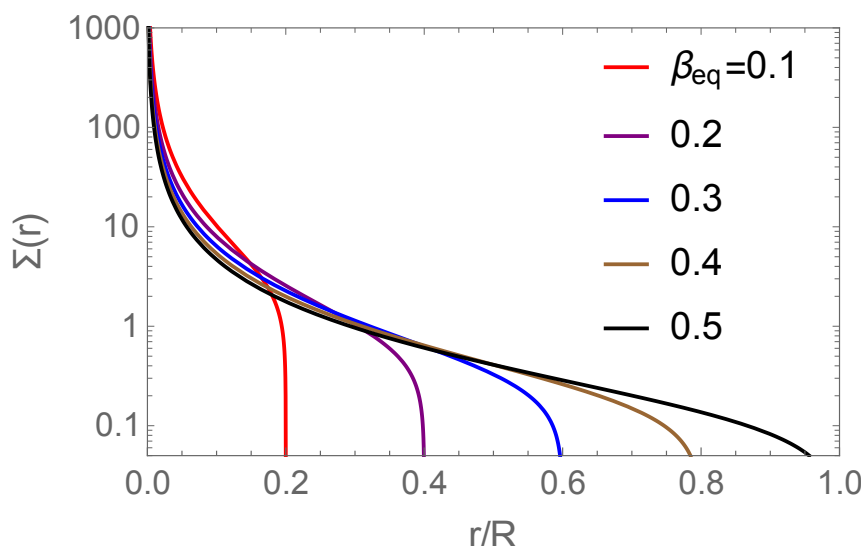


Figure 22.3 Disk surface density $\Sigma(r)$ (i.e., mass per unit area of the disk) vs. disk radius r for the simple model of the gravitational collapse of a rotating, initially spherical cloud. The curves show results for various initial equatorial ratios of the rotational to gravitational energy, $\beta_{eq} = 0.1 - 0.5$. The disk surface density here is in units of ρR , where ρ and R are the cloud's initial mass density and radius.

for the associated ratio $\beta(\theta)$ of the local rotational energy to gravitational energy, writing this in terms of the equatorial ratio β_{eq} derived in eqn. (22.12).

b. Use this to derive an expression for the associated disk radius $r(\theta)$ to which material contracts from various latitudes on the initial spherical surface of radius R . (You may assume that throughout the contraction, the gravitational attraction is that from a point source of mass M at the cloud center.) The blue lines in figure 22.2 draw connections between this disk radius and its source location at various latitudes on the cloud surface, for various choices of the parameter β_{eq} .

Challenge Problem: *Disk surface density*

a. Consider a *hollow* thin spherical shell of radius R and rotation frequency Ω that collapses under the gravitational attraction of a star of mass M_* at the shell center. Assuming the shell has a mass M_s that is initially spread uniformly over its spherical surface, use the results of the previous problem to derive an expression for the disk surface density $\Sigma(r)$ as a function of disk radius r . Express this in terms of the shell mass M_s , the outer disk radius r_d in (22.11), and the ratio r/r_d .

b. Now use this result to derive an integral expression for the total disk surface density $\Sigma(r)$ from collapse of a *filled*, constant-density, spherical cloud of radius R , mass M , and (rigid-body) rotation frequency Ω . (You may assume that the mass $M(r)$ inside any material initially at radius $r \leq R$ remains constant throughout the contraction.) Figure 22.3 plots results for such a disk model.

23 Origin of Planetary Systems

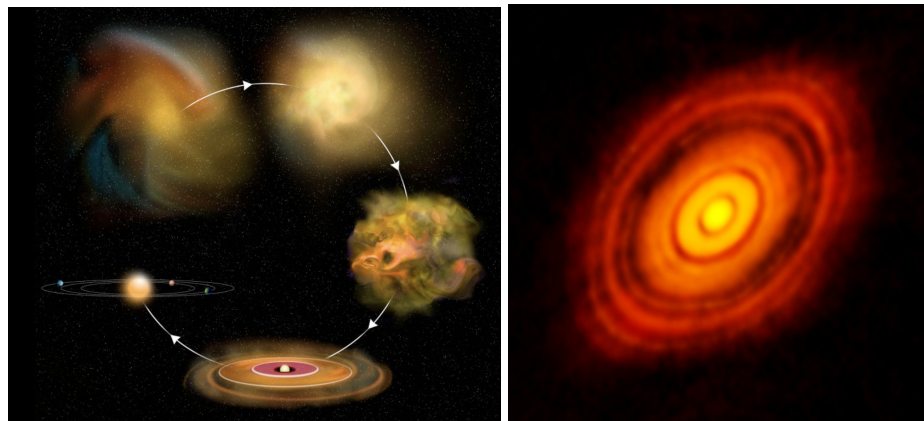


Figure 23.1 *Left:* Illustration of the nebular model for formation of a planetary system. *Right:* Direct image of protoplanetary disk in the T Tauri star HL Tauri, made in mm wavelengths with the Atacama Large Millimeter Array (ALMA). The entire disk spans about 200 au. The disk gaps likely represent regions of planet formation.

t

23.1 The Nebular Model

The disk formation process of the previous section forms the basis for the “Nebular Model” for the formation of planetary systems, including our own solar system. As illustrated in figure 23.1, as a protostellar cloud collapses under the pull of its own gravity, conservation of its initial angular momentum leads naturally to formation of an orbiting disk, which surrounds the central core mass that forms the developing star. With the usual interstellar composition of mostly Hydrogen and Helium, and only about 1-2% of heavier elements, this disk is initially gaseous, held in a vertical hydrostatic equilibrium about the disk mid-plane, with the radial support against stellar gravity provided by the centrifugal force of its orbital motion.

This stops the rapid, dynamical infall, but as the viscous coupling between dif-

ferentially rotating rings transports angular momentum outward, there remains a relatively *slow inward diffusion* of material that causes much of the initial disk mass to gradually accrete onto the young star. This, along with other effects – like the entrainment of disk material by an outflowing stellar wind – gradually depletes the Hydrogen and Helium gas in the disk. But during this slow dissipation of the disk mass, which likely occurs over a few million years, the heavier trace elements can, in the relatively dense conditions of this disk, gradually bond together to make molecules. These in turn nucleate into ever-growing grains of dust, and eventually into rocks and even boulders.

Collisions among these boulders leads to a combination of fragmentation and accumulation, with the latter eventually forming asteroid size (meters to kilometer) bodies for which self-gravity becomes significant. This first forms loosely held ‘rock piles’, then planetoids, and eventually planets. The detailed process are chaotic, with frequent collisions, but eventually, through accretion and assimilation the largest bodies clear out most of the debris that shared their orbital distance from the central star.

23.2 Observations of Protoplanetary Disks

While the basic ideas behind the nebular model date back to Kant and Laplace in the 18th century, modern observations now provide direct support for the overall model, and increasingly strong constraints on its specifics. Young stellar objects (YSOs), identified spectroscopically to still be contracting to the main sequence along the Hayashi or Henyey tracks (§17.4), often show clear evidence of protoplanetary disks. Herbig Ae/Be stars are relatively hot, massive YSO’s that show strong Hydrogen emission from a gaseous disks. The cooler, lower-mass T Tauri stars often show an infrared excess thought to arise from dust thermal emission in a warm protoplanetary disk.

With advent of telescope arrays (e.g. ALMA, see §13.2) observing in the far infrared and sub-mm spectral regions, it is now becoming possible to directly image such disks. The right panel of figure 23.1 shows an ALMA image of a protoplanetary disk in the T Tauri star HL Tauri, made in mm wavelengths. The star’s temperature and luminosity put it on the Hayashi track of the HR diagram, in a pre-main-sequence phase with age less than 1 Myr. Interferometry from the array allows spatial resolution ranging down to 0.025 arcsec. At HL Tauri’s distance of 140 pc, this corresponds to 3.5 au, with the visible disk extending over a diameter of about 200 au. The disk gaps likely represent regions where planet formation is clearing out disk debris, though there is so far no direct evidence of fully formed planets in this system.

Disks similar to that around HL Tauri have been inferred around other very young stars, but with densities that generally degrade with stellar age, over timescales of a few Myr. The Hubble space telescope has imaged several “debris disks” around stars with ages ~ 10 Myr. These are thought to be the later stages

when the disk debris has been depleted by various processes, like accretion onto the star, dissociation by stellar UV radiation, and entrainment in a outflowing stellar wind. A key issue in planet formation is thus whether this can occur quickly enough to compete with such disk depletion.

23.3 Our Solar System

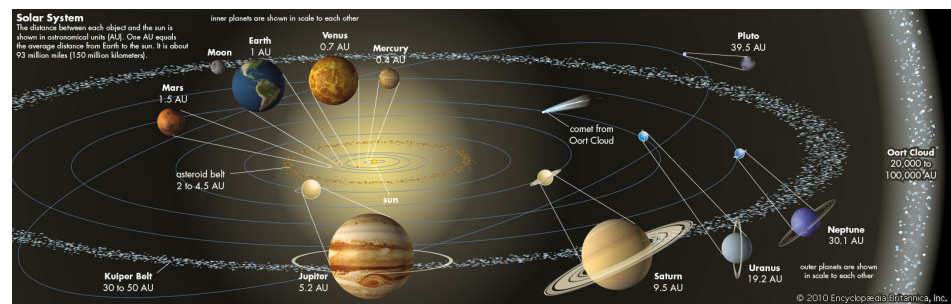


Figure 23.2 Illustration of key components of our solar system.

The above nebular model for formation of planetary systems implies that planets should be quite common around stars with less than a few solar masses. Indeed, §25 below discusses the techniques that have led to positive detection of some 4000+ such extra-solar planets (a.k.a. “exoplanets”). But our own solar system still provides a key, best-observed prototype, and so as background, let us first briefly review the key properties of the bodies and material that surround and orbit our Sun (see figure 23.2)

Chief among these are the 8 planets¹, which can be quite conveniently divided into the 4 relatively small, rocky inner planets (Mercury, Venus, Earth, Mars), and the 4 outer giants made of mostly gas (Jupiter and Saturn) and ice (Uranus and Neptune). They all have roughly circular orbits that lie in nearly the same plane as the orbit of the Earth, known as the *ecliptic*. Most have rotation that, with various tilts, are in the same sense as their orbit, the exceptions being Venus, which rotates slowly backward, and Uranus, whose rotational axis nearly lies in the plane of its orbit.

Between the orbits of Mars and Jupiter, there is a belt of smaller *asteroids* that likewise mostly have nearly circular orbits in this ecliptic plane. Most are small ($\sim 1\text{m}$ - 1km) with irregular shapes, but the largest, Ceres, with radius $R \approx 1200\text{ km}$, is massive enough to be made spherical by its self-gravity, and so is classified a “dwarf planet”. Asteroid orbits are strongly influenced by Jupiter’s

¹ The erstwhile ninth planet Pluto, has now be reclassified as a “dwarf planet”, and part of the Kuiper belt discussed below.

strong gravity, which is systematically clearing the region. Their combined mass is estimated to be only about 4% that of Earth's moon.

The gas giants all have multiple satellites, formed by a smaller-scale version of angular momentum conservation as their proto-planetary clouds were contracted by gravity. This left Jupiter and Saturn with several moons that, while much smaller than their host planet, have comparable size to Earth's moon. These include Jupiter's four 'Galilean' moons² (Io, Europa, Ganymede and Callisto), and Saturn's Titan, the only moon with a dense atmosphere, composed largely of nitrogen and methane. Saturn also has its prominent ring systems, which are composed of a large number of icy bodies ranging from centimeters to several meters across. The other gas giants have thinner, weaker rings. Jupiter's moon Europa also shows evidence for extensive surface ice, as well as a sub-surface ocean.

Somewhat beyond the orbit of Neptune are icy "Kuiper Belt Objects" (KBOs), which have somewhat eccentric and inclined orbits in a belt at distance 30-50 au. Originally predicted based on theoretical arguments, there are now more than 2000 directly detected, with more than 100,000 thought to exist with diameter >100 km. A few, including Pluto and its companion Charon, are large enough to also be considered dwarf planets. Indeed, their discovery was a major factor in the decision to reclassify Pluto as a KBO.

At much greater distances (>1000 au) lies the "Oort cloud" of icy planetesimals, with eccentric orbits that flare from near the ecliptic in the inner regions, to a nearly spherical distribution in the outer cloud. When deflected into the inner solar system, heating of the ice by the Sun causes outgassing, which with the outward push from radiation and the solar wind form the characteristic tail of comets.

23.4 The Ice Line: Gas Giants vs. Rocky Dwarfs

For the *Gas Giants* (Jupiter and Saturn) and *Ice Giants* (Uranus and Neptune) in the outer solar system, we need also to consider the role the much cooler conditions in allowing the formation of water ice.

As shown in figure 6.3, Hydrogen and Oxygen are the first and third most abundant elements in the Sun. In the solar nebula, their ready combination into water molecules (H₂O) thus made that relatively abundant. In the colder outer regions these condensed to form ice, which gradually collected into ever larger solid cores, eventually growing massive enough to gravitationally attract and retain the even-more-abundant but lighter gases of Hydrogen and Helium. This was the basis for formation of the outer gas giant planets, with an overall composition similar to the solar nebula, and the present-day Sun. In contrast, in the inner nebula, where it was too warm to form ice, such light atoms of

² So named because they were first discovered by Galileo.

Hydrogen and Helium escaped from the weaker gravity of the smaller, rocky planets, effectively stunting their growth and so keeping them relatively small.

To quantify this “ice line” between inner rocky dwarfs and outer gas giants, let us next derive an estimate for the decline of equilibrium temperature with distance from the Sun.

23.5 Equilibrium Temperature

For an absorbing sphere with radius r at a distance d from the Sun, the intercepted flux of the Sun’s luminosity L_\odot is $\pi r^2 L / 4\pi d^2 = \pi r^2 \sigma_{\text{sb}} T_\odot^4 (R_\odot / d)^2$, with R_\odot and T_\odot (≈ 5800 K) the Sun’s radius and effective temperature, and σ_{sb} the Stefan-Boltzmann constant (see §5.1). If we assume this sphere then radiates this energy as a blackbody over its surface area, $4\pi r^2 \sigma_{\text{sb}} T^4$, then solving for its equilibrium temperature gives

$$T(d) = T_\odot \sqrt{\frac{R_\odot}{2d}} \approx 290 \text{ K} \sqrt{\frac{\text{au}}{d}}. \quad (23.1)$$

Note here that the sphere’s radius r has cancelled, and so in principle this could be applied to bodies of any size, ranging from grains of dust to whole planets.

Indeed, for an object orbiting at the Earth’s distance of 1 au, the equilibrium temperature of about 290 K (i.e., just above water’s freezing point of 273 K) is pretty close to the actual mean temperature of Earth itself. But that is the result of a somewhat fortuitous and delicate cancellation, between the cooling effect of reflection of sunlight by clouds, and the warming effect of greenhouse gases in the Earth’s atmosphere.

By contrast, at a distance of about 0.7 au, the planet Venus would by eqn. (23.1) be predicted to have a temperature just about 15% higher than Earth, ~ 350 K; but in fact, due to a *runaway* greenhouse effect, the surface temperature on Venus is more than twice this value, > 700 K.

On the other side, for its distance at about 1.5 au, Mars has a predicted equilibrium temperature ~ 235 K. Owing to the lack of much greenhouse effect from its much thinner atmosphere, this is pretty close to Mars’ actual average surface temperature. While there is much evidence that Mars once had a much thicker atmosphere, and a warm enough temperature to have had liquid water flowing across its surface, today all its water is locked up in ice, at its poles and below its surface.

23.6 Questions and Exercises

Quick Question 1: a. Generalize the equilibrium temperature equation (23.1) to the case that a body has a non-zero *albedo*, $a > 0$, i.e. that it reflects a fraction a of incoming light, instead of absorbing it. b. Derive Earth’s equilibrium temperature given its mean albedo $a \approx 0.3$.

24 Water Planet Earth



Figure 24.1 Illustration of the Giant Impact model for formation of the Earth-Moon system.

24.1 Formation of Moon by Giant Impact

The large number of moons of the giant planets like Jupiter and Saturn likely formed through angular momentum conservation during the gravitational contraction of their protoplanetary gas cloud, effectively making them each mini-planetary systems on a smaller scale. The size and mass ratios of these moons to their host planets are very small, much like the ratio of planets to the Sun.

In this respect, the Earth's moon is quite distinct in being a comparable size ($\sim 1/4$) to Earth. Samples from the Apollo missions show the moon has an isotopic signature very similar to Earth, indicating it likely formed from material in the Earth's crust and mantle. Because it also lacks much of the iron that makes up the Earth's core, this led to the theory that even well after (perhaps 1-10 Myr) the proto-Earth had formed and the heavy iron had settled into its core, there was a *Giant Impact* by a third, Mars-sized body – often dubbed “Thea”. This impact ejected material from the Earth's mantle into orbit, which quickly cooled and condensed into the moon. (See figure 24.1.) Over billions of years, tidal coupling between the Earth and moon transferred angular momentum from the Earth's rotation to the moon's orbit, causing it to migrate from its initially close orbit to its present distance some 30 Earth diameters away.

While such a giant impact might seem rather unlikely and fine-tuned in the context of our present-day solar system, such events actually well represent the

chaotic conditions that reigned during the final phase of planets clearing out competing large bodies that overlapped with their own orbit.

24.2 Water from Icy Asteroids

Of course, the energy and heating of such an impact likely had major consequences in also expelling much of the volatile material – like water – that might have been retained from Earth’s initial formation. But the extensive cratering on the moon shows that even well after it formed, there were still ongoing extensive impacts from other bodies. This “Late Heavy Bombardment” is thought to have been triggered by ongoing gravitational interactions among the outer planets, leading to migration of Jupiter through the asteroid belt, and perhaps also a swapping of the orbital positions of Uranus and Neptune. This sent the icy minor bodies hurtling toward the inner solar system, to impact the moon, and of course also the Earth. In the vacuum on the moon, heating by the Sun melted and evaporated the volatile water, which was then lost into space¹; but on Earth, the ice from these ongoing impacts likely provided the source for much of the copious water that fills Earth’s oceans today. Comparing the isotopic signature of ocean water with recent analyses of space-mission samples from comets and icy asteroids indicates that the latter are most consistent with providing most of Earth’s water.

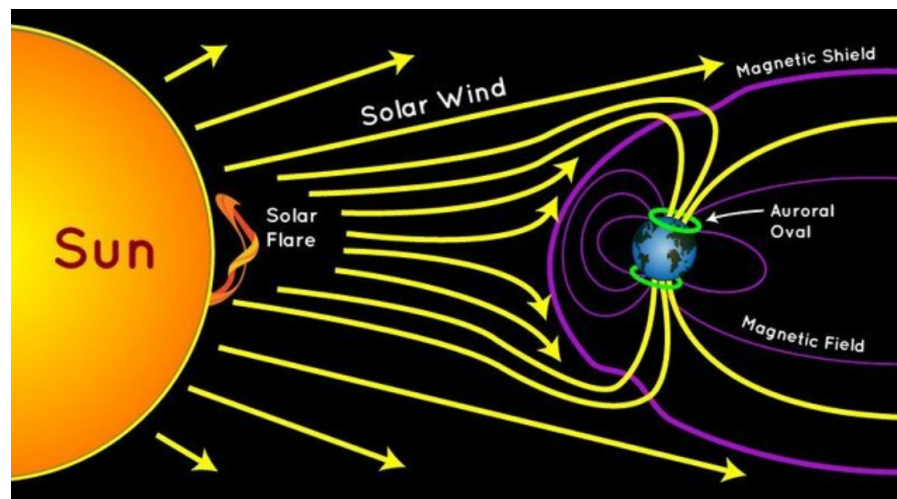


Figure 24.2 Illustration of how Earth’s magnetic field shields it from direct impact by the solar wind.

¹ except, evidence from lunar orbiters now suggests, in perpetually shadowed craters near the lunar poles, where still-frozen ice might prove a crucial resource for future lunar exploration

24.3 Our Magnetic Shield

Central to this retention of Earth's water is that fact that Earth has also retained a dense atmosphere, which then keeps the water cycling among oceans, air, and land. The gases in the Earth's atmosphere can be traced to outgassing from volcanoes, which themselves are a consequence of the plate tectonic collisions driven by escape of heat from radioactive decay in the Earth's mantle. Water acts like a lubricant for maintaining these tectonic movements.

By contrast, on Venus, the runaway greenhouse effect has effectively evaporated and dissociated its water, which then largely escaped into space. Without the lubricating effect of water, Venus' plates effectively stalled. Instead of being released gradually through tectonic activity, the internal heat of Venus thus builds up, then released in violent, planet-wide epochs of volcanic eruption every few hundred million years.

On the other side, the much smaller volume and mass of Mars implies its radioactive heating decayed away long ago, leaving now just a trace of extinct volcanoes. Moreover, the lack of a molten iron core meant there was no mechanism to produce the kind of global magnetic field that is generated by the convective dynamo in the Earth's core. Without such a global magnetic field, Mars is directly bombarded by high-speed protons from the solar wind, which over billions of years have been steadily eroding Mars' initial atmosphere, so that it now has only about 1% the surface pressure of Earth's atmosphere.

As illustrated in figure 24.2, the Earth's magnetic field effectively deflects these solar wind protons, forming a magnetospheric shield to protect its atmosphere (and us) from their direct impact. Instead it just guides some fraction of solar particles to impact near the magnetic poles, where they harmlessly light up the upper atmosphere to form the beautiful dance of the northern and southern lights (a.k.a. aurora).

24.4 Life from Oceans: Earth vs. Icy Moons

Life on Earth originated in these oceans some 3+ billion years ago, first as single cells that gradually collected into evermore complex, multi-cellular forms. In the fullness of time, some grew a spine and crawled onto dry land, eventually leading even to large land animals, including humans like us. But even our bodies and cells still retain about 60% water by mass, reflecting their ancient origins in the oceans. Water is the essential solvent that transports the nutrients of life.

In addition to water, life fundamentally requires an energy source. For most present-day life, this can be traced to the energy from the Sun, captured via photosynthesis by green plants.

But an exception to this lies in deep-ocean vents, where the energy of upwelling magma seeds formation of complex, energy-rich compounds (e.g., hydrogen-sulfide). These are then metabolized by giant worms and related organisms,

stoking a complex ecosystem that is largely isolated and independent of life near the surface or on land.

While the icy moons of gas giants are too cold to have liquid water on their surfaces, the tidal flexing that arises from their eccentric orbits around their host planet can generate enough internal frictional heat to warm an extensive subsurface ocean. This is thought to occur in at least two such icy moons of gas giants. Europa orbiting Jupiter shows a complex, mottled surface of ice quite similar to ice flows seen in Earth's arctic oceans. And the relatively small satellite Enceladus orbiting Saturn shows cracks near its south pole, from which water geysers have been directly observed, and indeed directly sampled by passages through them by the Cassini spacecraft.

While Cassini did detect some of the simplest chemical building blocks of life, its instruments were not designed to detect more complex molecules, or life itself. Plans are currently being developed to send further spacecraft to both these icy moons, with the aim to directly detect evidence for biochemistry or biological activity. If successful, this would for the first time extend our knowledge of life beyond our home planet Earth.

24.5 Questions and Exercises

Exercise 1:

- Given the moon's period $P = 28$ days and mean orbital distance $d_m \approx 400,000$ km, what is the moon's orbital speed v_m , in km/s?
- Given the moon's mass $M_m = 7.4 \times 10^{25}$ g, compute the moon's associated orbital angular momentum $J = M_m v_m d_m$ in CGS units.
- Approximating the Earth as a solid body with constant density, compute its moment of inertia $I = (2/5)M_e R_e^2$, given its mass $M_e \approx 6 \times 10^{27}$ g and radius $R_e \approx 6400$ km, again in CGS units.
- Next compute the Earth's rotational angular momentum $J_e = I\omega$ (in CGS units), where its angular rotation frequency $\omega = 2\pi/(24 \text{ hr})$.
- Suppose that the moon first formed at a distance of just $d_o = 2R_e$, compute its orbital speed v_o .
- Next, assuming the total angular momentum $J = J_e + J_m$ of the Earth-moon system is conserved, estimate the Earth's rotation period when the moon was at this distance.
- Finally, what is the Earth's associated equatorial rotation speed, and what fraction is that of the speed needed to reach near-surface orbit?

25 Extra-Solar Planets

25.1 Direct Imaging Method

Because they are much cooler than stars, the thermal emission from planets is mostly in the infrared. Their appearance at visible wavelengths comes instead by reflected light from their host star. This greatly complicates direct detection of extra-solar planets, since this reflected light is generally overwhelmed by the direct light from the star. Nowadays, there are some (~ 20) such direct imaging detections of exoplanets that appear far enough away from their host that it is possible to block out the light from the star without also blocking the planet.

Figure 25.1 shows the example of a sequence of 3 direct images, taken over 7 years by one of the Keck Observatory's two 10-m telescopes, of the 4 planets orbiting the star HR8799. The apparent shift in the positions of the planet images even allow one to infer their orbital periods, which range from 49 to 474 years.

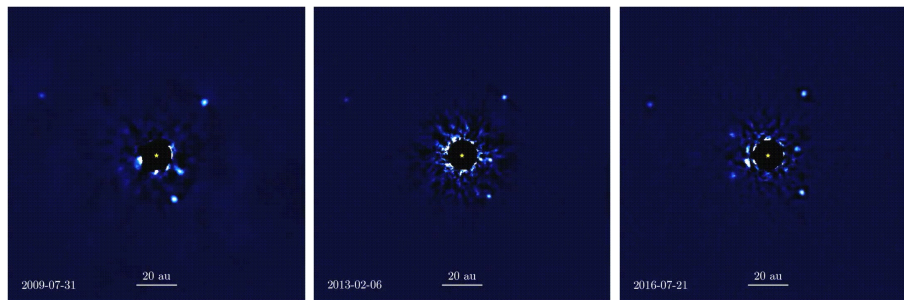


Figure 25.1 Sequence of 3 direct images by the Keck Observatory of the 4 exoplanets orbiting the star HR8799. The 3.5 year intervals (indicated by the dates labeled) show that the 4 planets are moving around the star, with inferred periods of 49, 100, 189, and 474 years. For reference, the orbital period of the Sun's outermost official planet, Neptune, is 165 years. The planets are visible because most of the light from the star is blocked out by an occulting disk. The 3 planets to the right of the star are clearly seen; a fourth, much dimmer one, can be found to the upper left. The lower bar showing the extent of 20 au indicates the planets have orbital distance of several tens of an au.

But most exoplanet detections have been made via two other more indirect

techniques, known as the *radial velocity* and *transit* methods, as illustrated by figure 25.2.

Each of these three methods have analogs in the study of stellar binary systems, as outlined in section 10. Direct imaging is similar to visual binary systems (section 10.1); the radial velocity method is similar to spectroscopic binaries (section 10.2); and the transit method is analogous to eclipsing binaries (section 10.3).

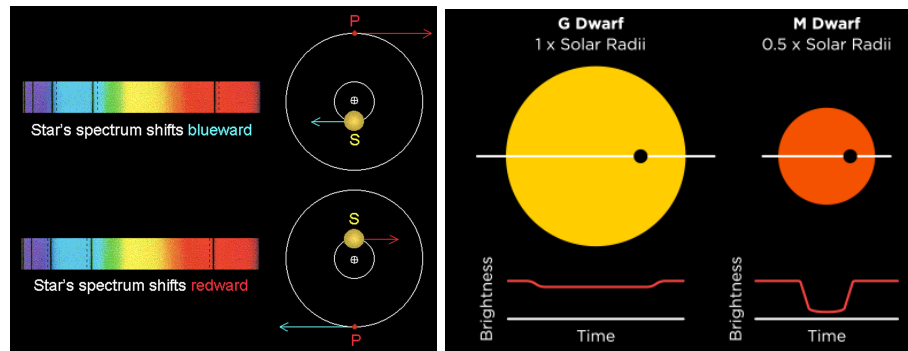


Figure 25.2 *Left:* Illustration of the radial velocity method, showing how the orbit of a planet P causes a wobble of its star S; as the star alternatively moves toward or away from the observer (here to the left), this leads to small blueward or redward Doppler shifts of spectral lines in the star’s spectrum. Note, the *dashed* lines here represent the original, *rest wavelength* positions of the spectral lines. *Right:* Illustration of the transit method, showing how the light from a star slightly dims when the dark planet goes in front of the star; the net effect is stronger for a cooler, smaller red dwarf star (spectral type M) than for a solar-type G dwarf with higher temperature and larger radius.

25.2 Radial Velocity Method

As illustrated in the left panel of figure 25.2, the radial-velocity method refers to the periodic movement of the host star due to the gravitational pull of the planet. The associated spatial “wobble” is not directly detectable, but as in spectroscopic binaries, its associated motion toward and away from the observer can be detected via very precise spectroscopic measurements of the systematic Doppler shift from multiple absorption lines in the star’s spectrum.

Referring to equation (10.8) from our discussion of spectroscopic binary systems, let us identify object 1 with the planet of mass $M_1 = m_p$ and object 2 with the star that has an orbital speed $V_2 = V_*$. Using the fact that $V_1/V_2 =$

$M_2/M_1 = M_*/m_p$, we find for the star's speed

$$\begin{aligned} V_* &= \left(\frac{2\pi G M_{tot}}{P} \right)^{1/3} \frac{m_p}{M_{tot}} \\ &\approx \left(\frac{2\pi G M_*}{P} \right)^{1/3} \frac{m_p}{M_*} \\ &\approx 30 \text{ km/s} \left(\frac{M_*/M_\odot}{P/\text{yr}} \right)^{1/3} \frac{m_p}{M_*}, \end{aligned} \quad (25.1)$$

where the second equality makes use of the fact that the planet's mass is negligible compared to the star, $M_{tot} = M_* + m_p \approx M_*$. This shows that the wobble speed is directly proportional to the planet's mass, but scales with the inverse cube root of the orbital period. The former favors planets with bigger mass, while the latter favors planets that orbit close to their parent star.

The very first detection of an exoplanet around another star was by this radial velocity method. It indicated what is now known as a *Hot Jupiter*, i.e., a planet with a mass comparable to (actually even larger than) Jupiter, but orbiting at such a close distance that the stellar heating would make it quite hot. In the context of the prevailing idea (discussed in section 23.4) that such gas giants should only form beyond the ice line, this detection was a real surprise. It led to a revised view that, while such Hot Jupiters were indeed formed in the outer regions beyond the ice line, gravitational interaction with other giant planets out there led some to be flung into an *inward migration*, so that they finally ended up very close to their star.

Even if such complex interactions and migrations are rare, just the fact that they can occur at all can explain initial prominence of Hot Jupiter detections, which according to eqn. (25.1) are, after all, observationally favored. This illustrates that, to assess the relative importance of such an observed phenomenon, it is important to be aware of the inherent *observational biases* that come with a given method or technique. It also means that it is helpful to identify other independent, and hopefully complementary, techniques.

For lower-mass planets orbiting with longer periods further from the star, the wobble velocity is smaller, and so can be hard to detect. For example, for the Earth orbiting the Sun, we have $P = 1 \text{ yr}$ and $M_e/M_\odot \approx 3 \times 10^{-6}$, giving $V_* \approx 9 \text{ cm/s}$. With current technology, the smallest measurable speeds are about a factor ten times higher; but because of the obvious interest in detecting Earth-size planets orbiting at a distance of 1 au around a Sun-like star, being able to detect wobble speeds near this Earth-Sun value is an ultimate, long-term goal.

25.3 Transit Method

Fortunately, there is another, quite-distinct way to detect exoplanets, known as the *transit method*. As illustrated in the right panel of figure 25.2, this simply

looks for the slight dimming of the star's apparent brightness whenever a planet "transits" in front of it. Instead of elaborate spectroscopic measurement of the slight Doppler shift, this merely requires precise photometric measurements of changes in the star's total apparent brightness. It is essentially analogous to eclipsing binaries discussed in section 10.3 and illustrated in figure 10.3, except that the lower temperature star is now just replaced with a planet. Due to its even lower temperature, the planet emits mainly in the infrared, with little intrinsic emission in the visible. Thus the fractional drop in the star's observed apparent brightness is just set by the ratio of the projected areas of the planet vs. star, which scales with the square of the ratio of their radii,

$$\frac{\Delta F}{F} = \left(\frac{R_p}{R_*} \right)^2. \quad (25.2)$$

As illustrated in figure 25.2, the net change is thus greater for a smaller, red dwarf (type M) star than for a larger, Sun-like (type G) star.

The minimum size planet that can be detected depends mainly on how precisely one can measure the stellar brightness. For ground-based telescopes, the main source of noise comes from distortions and variations from the Earth's atmosphere. The minimum planet-to-star size-ratio scales with the square root of that noise. For example, a typical noise level of 1% allows one to detect a Jupiter size planet, with $R_p/R_* \approx 0.1$. That is readily achieved even with ground-based telescopes.

But detecting smaller, rocky planets like Earth, which have $R_p/R_* \approx 0.01$, requires a factor 1/100 lower noise, i.e. about 0.01%. This generally requires telescopes in space.

Another factor for the transit method is that it only works for planets whose orbital planes are near our line of sight. For a planet of radius r at a distance d from a star with radius R , the angle α that the line of sight makes to the planet's orbital must be within

$$\alpha_{\max} = \arctan \left(\frac{R+r}{d-R} \right) \approx \frac{R}{d} = 0.0047 \frac{R/R_\odot}{d/\text{au}} = 0.27^\circ \frac{R/R_\odot}{d/\text{au}}. \quad (25.3)$$

The latter equalities show that detecting transit an Earth-size planet around a solar-type star, the alignment must be within an angle $\pm 0.27^\circ$. Over the full angle range from zero to 90° , the associated probability of finding such Earth-like planet is only $P_e = 0.27/90 \approx 0.003$, or only 3 in every 1000.

Thus one generally has to monitor quite precisely a large number of stars to find those few that have this fortuitous alignment. A great breakthrough for this came from the **Kepler** satellite mission, named after the famous scientist who discovered the laws of planetary motion. Its prime mission was to monitor simultaneously the brightness of about 150,000 stars for several years, with a cadence of 30 minutes (and even every two minutes for a subsample). In addition to discovering planets, it also detected a wide range of phenomena associated with stellar brightness variations, like starspots that modulate brightness with

the star's rotation, or stellar pulsations. To discriminate planet transits from these other causes of variability, *Kepler* monitored stars long enough to capture repeat transits over several orbital periods, sometimes ranging up to year or more.

The probability of seeing a transit is highest for planets close to the star, which also tend to have the shorter, and thus more repeatable, periods. Thus most of the confirmed planets tend indeed to be from close orbits, and with large radii. But with extended monitoring over many years, it has become possible to identify a few planets with sizes near that of Earth, orbiting at distances up to an au.

Unfortunately, degradation of *Kepler*'s guiding gyroscopes eventually forced the so-called K2 phase of the mission to focus on different patches of sky, and finally to discontinue operations altogether. A follow-up mission called TESS (Transiting Exoplanet Survey Satellite) is systematically mapping the full sky, monitoring stellar photometric variability in the search for transiting planets.

25.4 The Exoplanet Census: 4000+ and counting

As of this writing (Fall 2020), there have been 4000+ confirmed exoplanets, with several new ones added every day¹. Figure 25.3 plots the overall population in terms of planet size (relative to Earth) and planet orbital period. The legend for point style or color shows the discovery method, with the majority detected by the *Kepler* mission, based on the transit method. The radial-velocity method is quite successful at detecting both Hot Jupiters and Cold Gas Giants, but there are only a handful of planets found by direct imaging. Pulsar timing was actually the method for the very first detection ever, but it only applies in the rare case of a planet around a pulsar. A new method using gravitational micro-lensing shows promise for detecting relatively small planets far from their host star.

Figure 25.3 also outlines the various distinct classes of planets. While early discoveries were dominated by Hot Jupiters and Cold Gas Giants, transit surveys like *Kepler* have now identified numerous smaller planets. These range from the intermediate size Ice Giants, which when closer to their star become “Ocean Worlds”, to Rocky Planets, which when very close can become “Lava Worlds”, as the rocks are melted by intense heating from the star's radiation.

Note in fact the large number of planets detected with periods ranging from tens of days to even less than a day, implying they orbit much closer to their star than Mercury, with an orbital period of 88 days, is to our own Sun (0.4 au).

Such a more extensive survey with a variety of methods has given us a better understanding of observational biases. This allows one to compensate for the early dominance of Gas Giants and Hot Jupiters, which now are understood to

¹ A running compilation is given in https://exoplanetarchive.ipac.caltech.edu/docs/counts_detail.html

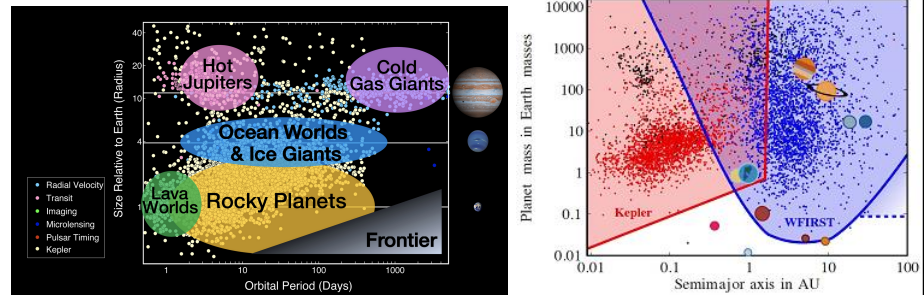


Figure 25.3 *Left:* Compilation of detected exoplanet populations, plotting planet size relative to Earth vs. inferred orbital period. The color key gives the discovery method. *Right:* Known planets discovered by Kepler (red) and all other methods (black), on a log-log plot of planet mass vs. semi-major axis of orbit. The blue dots compare the predicted detections by the planned WFIRST satellite, using gravitational micro-lensing. Note that this method will favor detection of planets at much larger distances from their host star. The planet icons represent masses and orbital distances of planets in our solar system.

be relatively rare. Indeed, the most numerous class are rocky planets somewhat larger than Earth, so-called “Super-Earths”.

25.5 Search for Earth-sized Planets in the Habitable Zone

A key goal for ongoing exoplanet searches is to detect Earth-size planets in the “Habitable Zone”, generally defined to be where liquid water could exist on the planet’s surface. As noted in our discussion in section 24 of our own Earth, the surface temperature can be affected by both the planet’s reflection of visible light (e.g. from clouds), and the greenhouse trapping by the atmosphere of cooling radiation in the infra-red. Since these are difficult to determine and quantify, a first approach is to assume that, as on Earth, these two effects roughly cancel, and so use just the simple blackbody equilibrium form derived in section 23.5. For a star with surface temperature T_* , eqn. (23.1) can be generalized to

$$T(d) = 290 \text{ K} \left(\frac{T_*}{T_\odot} \right) \sqrt{\frac{\text{au}}{d}} = 290 \text{ K} \left(\frac{T_*}{T_\odot} \right) \left(\frac{M_\odot}{M_*} \right)^{1/6} \left(\frac{\text{yr}}{P} \right)^{1/3}, \quad (25.4)$$

where the latter equality uses Kepler’s 3rd law ($M \sim d^3/P^2$) to obtain a scaling with orbital period P , while accounting for a relatively weak additional dependence on stellar mass M_* .

For a lower-mass star with also a lower surface temperature, the period to have an equilibrium temperature the same as Earth is thus

$$\frac{P_e}{\text{yr}} = \left(\frac{T_*}{T_\odot} \right)^3 \left(\frac{M_\odot}{M_*} \right)^{1/2}. \quad (25.5)$$

For example, for a red-dwarf star with $M_*/M_\odot = T_*/T_\odot = 1/2$, we find $P_e = 65$ day, with a corresponding distance $d_e = 0.25$ au.

Such close-in, potentially habitable planets around cool, low-mass, red-dwarf stars are easier to detect, both directly and by the radial-velocity and transit methods; so there are already quite a few such candidates. However, because such cooler stars have deeper convection zones, they tend also to show quite extensive magnetic activity, with flares and coronal mass ejections. An ongoing area of study, known as “Living with a Star”, seeks to examine how viable life could be affected on such close-in planets to active red dwarfs.

Another goal is to use subtle details of the observed spectrum from a star undergoing a transit to try to infer information on the planets atmosphere, e.g., from absorption or the starlight by molecules in the planetary atmosphere that would not exist on the much hotter star. In particular, any signature of molecular oxygen would be viewed as an indicator for life, since this is normally very reactive and would be destroyed unless constantly being replenished by photosynthesis.

25.6 Questions and Exercises

Quick Question 1: If we approximate the outer and inner limits of a habitable zone as ranging from where the equilibrium temperature is in the range $0\text{ C} < T < 50\text{ C}$, derive expressions, analogous to eqns. (25.4) and (25.5), for the inner and outer values for the orbital period P_i and P_o , and for the orbital distances d_i and d_o .

Exercise 1: *Transit and radial velocity signatures of Jupiter and Earth*

- If a Jupiter-size planet transits a solar-size star, by about what fraction is the star’s light reduced?
- If an Earth-size planet transits a solar-size star, by about what fraction is the star’s light reduced?
- If a Jupiter-mass planet orbits a solar-mass star with a period of 1 year, about what is the star’s wobble speed?
- If a Earth-mass planet orbits a solar-mass star with a period of 1 year, about what is the star’s wobble speed?
- A planet orbits a star that is twice as hot as the sun at a distance of 4 au. About what is the planet’s equilibrium temperature?

Exercise 2: *Radial velocity detection of habitable planets*

- From eqn. (24.1) compute the amplitude of the Sun’s wobble speed (in m/s) due to Jupiter’s orbit. (You’ll need to look up Jupiter’s orbital period and its mass ratio to the Sun.)
- Ignoring Jupiter and other planets, similarly compute the amplitude of the Sun’s wobble speed (in m/s) due to the Earth’s orbit.
- The smallest wobble speed that can be currently measured in a star’s spectrum is about 1 m/s. What does this imply about our ability to detect analogs of Jupiter and Earth around stars with mass comparable to our Sun.
- Next combine eqns. (24.1) and (24.3) to derive an expression for the wobble speed of a star with mass M_* and temperature T_* due to a habitable zone planet of mass m_p and orbital period P_e .

- e. For a star with a mass and temperature that are both $1/2$ that of the Sun, estimate the smallest mass planet m_p (in units of Earth's mass m_e) that could be detected in this star's habitable zone.
- f. How far (in au) would such a planet be from its star?

Part IV

Our Milky Way & Other Galaxies

26 Our Milky Way Galaxy

26.1 Disk, halo, and bulge components of the Milky Way

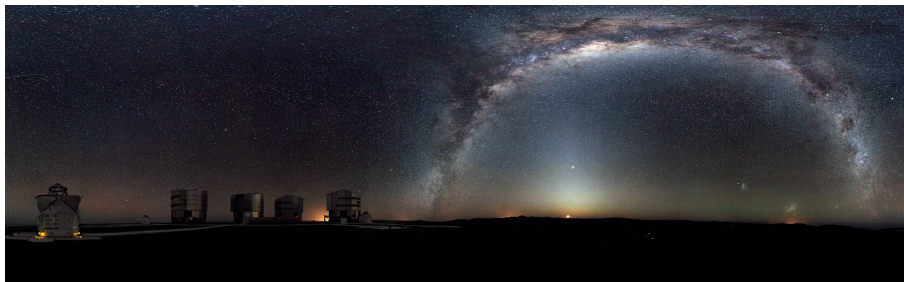


Figure 26.1 Panoramic photo of the Milky Way, taken from the European Southern Observatory’s facility in Paranal, Chile, located in the dry, and isolated, Atacama desert. The left side shows silhouettes of four 8m telescopes, along with a smaller 1.8m telescope in the left foreground. On the far right near the horizon the two hazy patches are dwarf galaxies that are satellites of our own Milky Way, the Large and Small Magellanic clouds. Though they appear to hang side-by-side, they are actually separated by about 15 kpc, and are removed from us by distances of 50 kpc and 60 kpc respectively.

The tendency for conservation of angular momentum of a gravitationally collapsing cloud to form a disk (see §22.5) is actually a quite general process that can occur on a wide range of scales: from planets, to proto-stellar cores, to even an entire proto-galaxy, with hundreds of billions times the mass of individual stars. This indeed provides the basic rationale for the disk in our own Milky Way (MW) galaxy. We along with our Sun are today still embedded within the Milky Way’s disk, orbiting about the galactic center, again because our bits of proto-galactic matter had too much angular momentum to fall further inward.

As we look up into a dark night sky, we can trace clearly the direction along this disk plane through the faint *milky glow* of thousands of distant, unresolved stars, from which we indeed get the name “Milky Way”. Figure 26.1 gives a vivid illustration of the Milky Way through a panoramic image of the night sky seen from the exceptionally dark and clear site of the Paranal observatory, in the Atacama desert of Chile. Indeed, toward the horizon on the right one can

also see two satellite galaxies of the Milky Way, known as the Large and Small Magellanic¹ Clouds (LMC and SMC).

As we look along this disk plane of the MW, the background/foreground superposition of many stars and GMCs makes it very difficult to discern the overall structure, the way we readily can from the face-on view of M51 in figure 21.5. Moreover, the extinction from the extensive gas and dust means that visible images, like those in figure 26.1 or in panel e of figure 21.2, only penetrate a limited distance, typically ~ 1 kpc, within the disk, which itself is only about 1000 ly, or just 0.3 kpc, in thickness.

Fortunately, IR and radio images can penetrate much further, even to the other side of the galaxy, spanning the full 100,000 ly (~ 30 kpc) diameter of this disk. Thus with painstaking work applying various methods for determining the distance to the myriad of stars and GMCs detected, it has become possible to draw a quite complete map of the overall disk structure of our MW galaxy, as given in figure 26.2. This shows that, like M51, our galaxy also has distinct spiral arms, along which are concentrations of gas, dust, HII regions, GMCs, and active star formation. The map nicely illustrates the position of our Sun well away from galactic center, and also serves to define the Sun-centered galactic longitude system used to chart the galactic disk.

Quick Question 1: *Galactic Year*

At the distance $d \approx 8$ kpc of the Galactic Center, the Sun turns out to have an orbital speed $V_o \approx 220$ km/s. How long is one “galactic year”, i.e., the Sun’s orbital period (in Myr) around the galaxy?

Figure 26.3 illustrates schematically the overall 3D morphology of the MW, which in addition to the *disk*, has distinct “*halo*” and central “*bulge*” components.

The halo is roughly spherical, with a diameter comparable to that of the disk, about 30 kpc. It contains very little gas or dust, and without much source for new star formation, its stars (dubbed “Population II”) are very old. This can be seen from the H-R diagrams of the *globular clusters* that are common in the halo, which typically have main-sequence turnoff points below the luminosity of the Sun, implying ages $t > t_{ms,\odot} \approx 10$ Gyr. These old globular clusters contain of order $10^4 - 10^5$ stars, and are much more gravitationally bound and stable than the “galactic” or “open” clusters found in the disk.

Such open clusters are typically quite young, with main sequences that sometimes extend to masses of many tens of solar masses, implying ages less than their main sequence lifetimes, i.e. less than a few times 10 Myr. They are irregular in shape, and typically only contain 100 or so stars. They are so loosely bound that they tend to disperse within a few 10 Myr or less, evolving into unbound *OB associations*. Due to tidal effects from the galaxy, along with the shear from its differential rotation, the stars eventually disperse and mix with other stars

¹ “Magellanic” because they were first reported to European civilization by Ferdinand Magellan, following his first-in-history circumnavigation of the Earth, with routes around southern continents showing the southern sky where these clouds are visible.

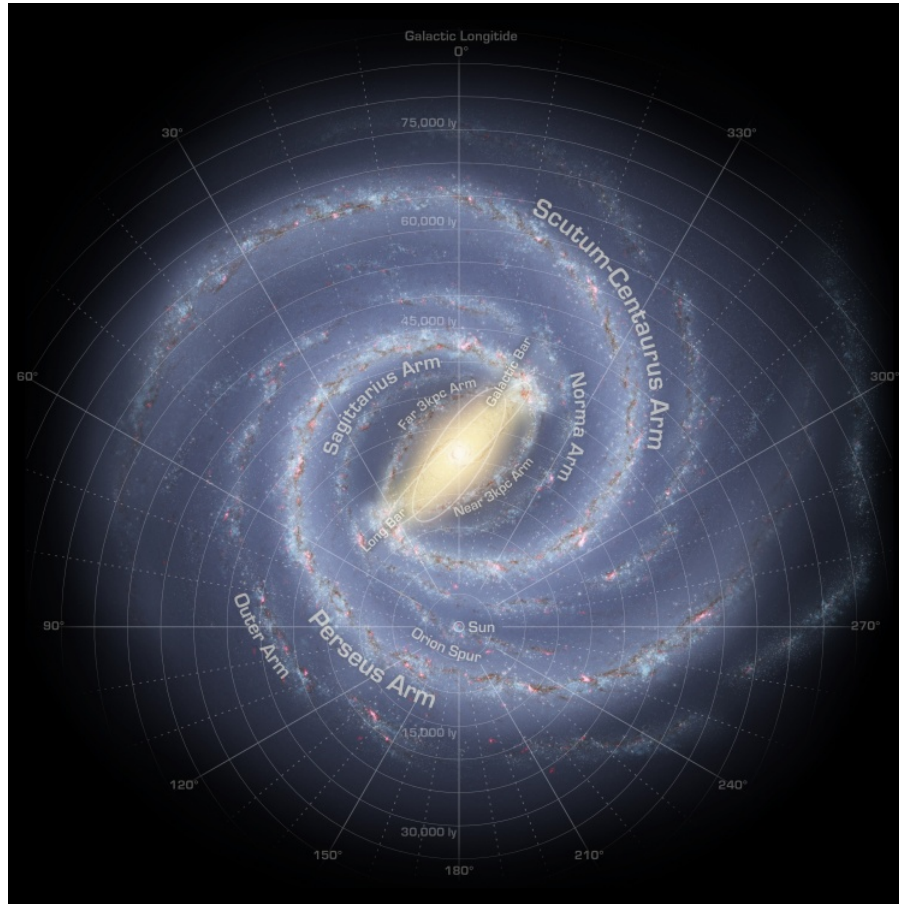


Figure 26.2 Map of the disk plane of our Milky Way galaxy, based on IR and radio surveys. Our Sun lies along the Orion spur of the Sagittarius spiral arm, between the inner Scutum-Centaurus arm, and the outer Perseus arm. The Galactic Center (GC), toward the constellation Sagittarius, is defined to be at zero galactic longitude, with the Sun orbiting a distance $d \approx 8$ kpc from the GC, in the direction of longitude 90° (i.e., to the left in the picture), toward the bright star Vega in the constellation Hercules. (Credit:modification of work by NASA/JPL-Caltech/R. Hurt (SSC/Caltech)).

(called “Population I”) in the disk. Figure 26.4 compares examples of a globular and an open cluster, and gives a Venn diagram showing the common and distinct properties between the two types.

The central bulge contains a mixture of traits of both the disk and halo, with both types of clusters, and both populations of stars (I and II). Because of dust absorption from within the galactic disk, it doesn’t appear in the visible to be much brighter than higher galactic longitude regions away from the galactic

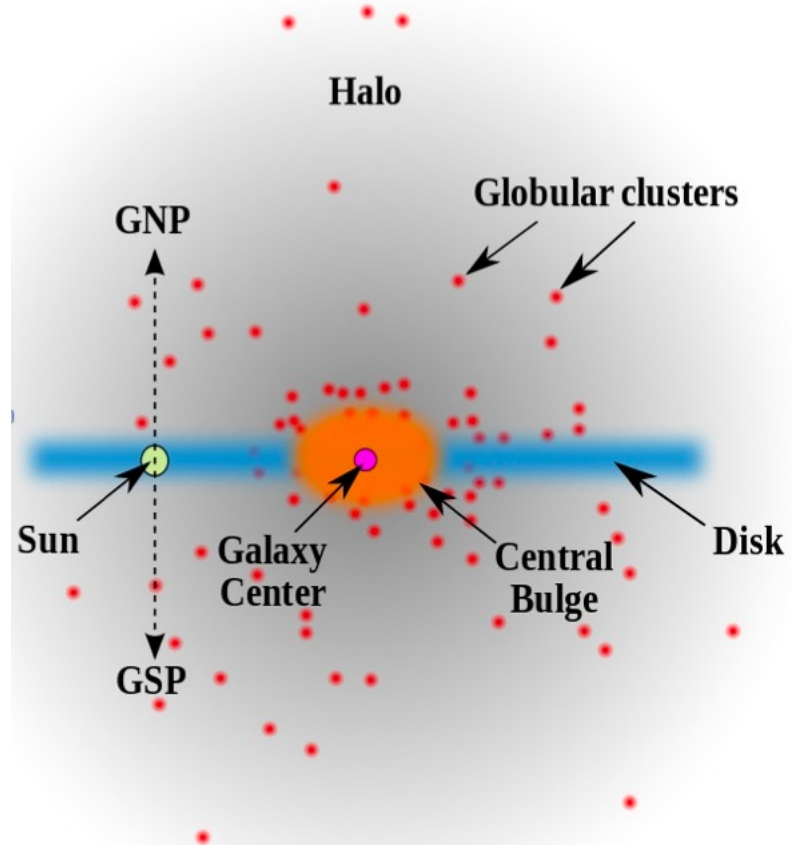


Figure 26.3 Edge-on schematic illustration of the 3D morphology of the MW galaxy, showing the disk, halo, and bulge components, with globular clusters in halo, and the Sun in the disk, offset from the galactic center. The directions from the Sun away from the disk plane are dubbed the Galactic North and South Poles (GNP and GSP).

center; but if one corrects for this absorption, it dominates the overall galactic luminosity (as can be seen from the case of M51 shown in figure 21.5).

26.2 Virial mass for cluster from stellar velocity dispersion inferred from Doppler shifts

By measuring the Doppler shifts of spectral lines from stars in an open cluster or a globular cluster, one can determine each star's radial velocity V_r . We can use this to define an average cluster radial velocity, $V_c \equiv \langle V_r \rangle$, as well as an *root mean square* (rms) *velocity dispersion* about this mean,

$$\sigma_v \equiv \sqrt{\langle (V_r - V_c)^2 \rangle}. \quad (26.1)$$

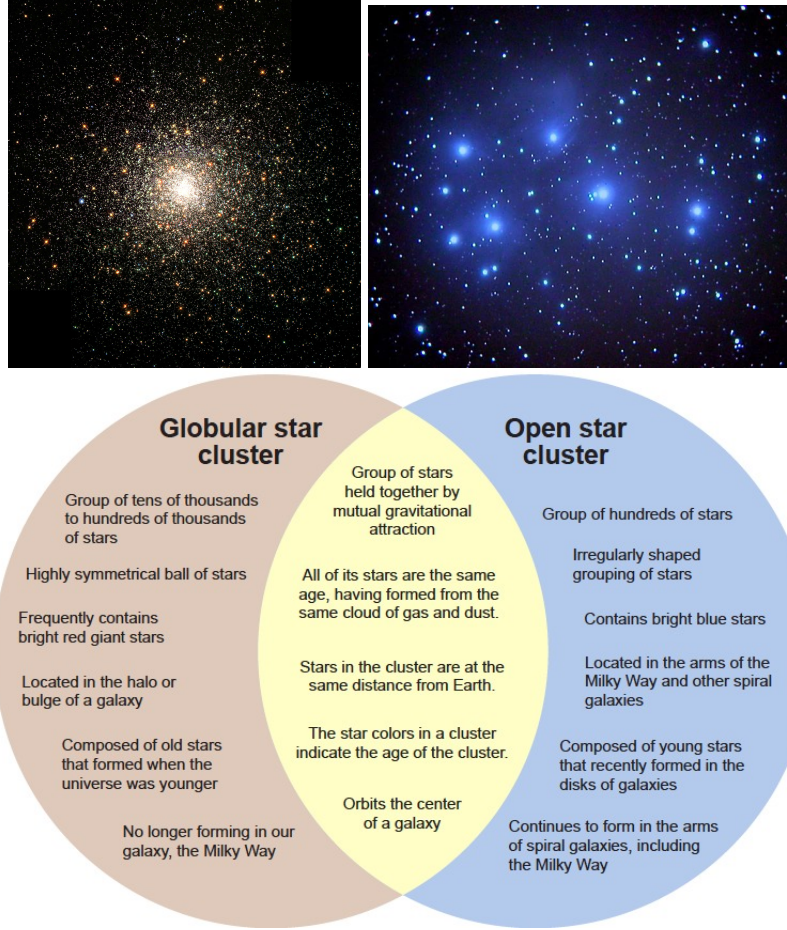


Figure 26.4 *Left:* The globular cluster M80. *Right:* The open cluster M45, a.k.a. the Pleiades or Seven Sisters. *Bottom:* A ‘Venn diagram’ comparing the different and common characteristics of Globular vs. Open clusters.

The kinetic energy-per-unit-mass associated with this random component of radial velocity dispersion is $\sigma_v^2/2$. Assuming a similar dispersion in the two transverse directions that cannot be measured from a Doppler shift, the total associated kinetic energy from the 3 directions of motion is $K = (3/2)M_c\sigma_v^2$, where M_c is the total stellar mass. For a cluster of radius R , the associated gravitational binding energy scales as $U \approx -GM_c^2/R$. If the cluster is bound, then application of the usual virial condition $K = |U|/2$ allows one to obtain the cluster mass via

$$M_c = \frac{3\sigma_v^2 R}{G} = 6.9 \times 10^4 M_\odot \left(\frac{\sigma_v}{10 \text{ km/s}} \right)^2 \frac{R}{\text{pc}}. \quad (26.2)$$

In practice, application of this method requires we obtain the cluster radius through the measured angular radius α and an independently known distance d , through the usual relation $R = \alpha d$.

26.3 Galactic rotation curve & dark matter

As illustrated in figure 21.2a, a primary diagnostic of atomic Hydrogen in the disk plane of the galaxy comes from its radio emission line² at a wavelength of $\lambda = 21$ cm. As we peer into the inner disk regions of the galaxy, i.e. along galactic longitudes in the range $-90^\circ < \ell < 90^\circ$, we find that this 21 cm line shows a distinct wavelength broadening $\Delta\lambda(\ell)$ that varies systematically with the longitude ℓ . Most of this broadening arises from cumulative Doppler shift along the line of sight from the motion associated with the orbit of distinct gas clouds about the galactic center; it thus provides a key diagnostic for determining the galaxy's "rotation curve" as a function of galactic radius R .

Quick Question 2: *Energy of Hydrogen 21-cm transition.*

What is the energy E , in eV, of the hyperfine, spin-flip transition that gives rise to the 21-cm emission line of neutral Hydrogen.

Figure 26.5 illustrates the basic geometry and associated trigonometric formulae. Focusing for convenience on longitudes in the range $0 < \ell < 90^\circ$, we find that the broadening actually takes the form of a redshift to maximum wavelength λ_{max} , which occurs when the line of sight along that longitude ℓ is *tangent* to some inner radius, $R = R_o \sin \ell$, where R_o (≈ 8 kpc) is the radius of our own galactic orbit along with the Sun. Thus by measuring λ_{max} , we can readily infer the *maximum* line-of-sight velocity away from us, $V_{rmax} = c(\lambda_{max}/\lambda - 1)$. Because our line of sight along ℓ is tangent to an orbit at this radius, the inferred maximum velocity just depends on the difference between the orbital velocity at R and the projection of our own orbital motion along this direction,

$$V_{rmax}(\ell) = V(R) - V_o \sin \ell = \Omega(R)R - \Omega_o R_o \sin \ell = (\Omega(R) - \Omega_o) R_o \sin \ell, \quad (26.3)$$

where Ω_o and $V_o = \Omega_o R_o$ are the angular and spatial velocity of the Sun's orbit at radius R_o . This can readily be solved to give the galactic rotation curve in terms of either the spatial or angular velocity

$$\Omega(R_o \sin \ell) = \frac{V_{rmax}(\ell)}{R_o \sin \ell} + \Omega_o ; \quad \boxed{V(R_o \sin \ell) = V_{rmax}(\ell) + V_o \sin \ell}. \quad (26.4)$$

The Sun orbits the galaxy at a radial distance $R_o \approx 8$ kpc from the galactic center, with a speed $V_o \approx 220$ km/s, implying then an orbital period $P_o \approx 220$ Myr. (See QQ1.)

² This results from a "hyperfine" transition in which the spin of the electron goes from being parallel to anti-parallel to the spin of the proton. The energy difference is much smaller than for transitions between principal energy levels of the Hydrogen atom, which are a few eV, and so have wavelengths of a few hundred nm, in the visible or UV spectral bands. See QQ 2.

Galactic Rotation

$$\begin{aligned}
 c \left(\frac{\lambda_{max}}{\lambda_0} - 1 \right) &= V_{max} \\
 &= V - V_o \sin \ell = \Omega R - \Omega_o R_o \sin \ell \\
 &= (\Omega - \Omega_o) R_o \sin \ell
 \end{aligned}$$

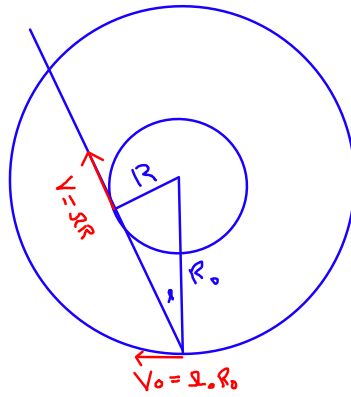


Figure 26.5 Sketch to show how measuring the change with galactic longitude ℓ of the maximum Doppler-shifted wavelength λ_{max} of the $\lambda_o \approx 21.1$ cm line from atomic Hydrogen can be used to determine the galactic rotation rate $\Omega(R)$ as a function of radius R , given the known rotation speed $V_o = \Omega_o R_o \approx 220$ km/s at the radius R_o of the Sun's orbit. The results indicate that, inside the Sun's orbit ($R \leq R_o$), the rotation speed is nearly constant, with $V(R) \equiv R\Omega(R) \approx V_o$.

Applications of this approach to analyzing observations of the 21 cm line of atomic H yield the rather *surprising* result that, within most of the region within the Sun's galactic orbit, $R < R_o$, the orbital speed is nearly *same* as that of the Sun, i.e.

$$V(R) \approx V_o \approx 220 \text{ km/s} ; \quad R < R_o, \quad (26.5)$$

which is known as a “flat” rotation curve.

Extension of this 21-cm method to longitudes $90 < \ell < 270$ that point *outward* to larger galactic radii $R > R_o$ is complicated by the need now to have an independent estimate of the distance to an observed Hydrogen cloud. But when this is done, the results indicate that the rotation curve remains nearly “flat”, with *constant orbital speed*, out to the farthest measurable radii, $R \lesssim 15$ kpc. The left panel of figure 26.6 compares this observed flat rotation curve for our galaxy vs. what would be expected from Kepler's law if the galaxy's mass were as strongly centrally concentrated as its stellar luminosity.

This comparison illustrates why these flat rotation curves came as a surprise. Since most of galaxy's *luminosity* comes from the central bulge within a radius $R_{bulge} \approx 1$ kpc, it seemed reasonable to presume that most of the galaxy's *mass*

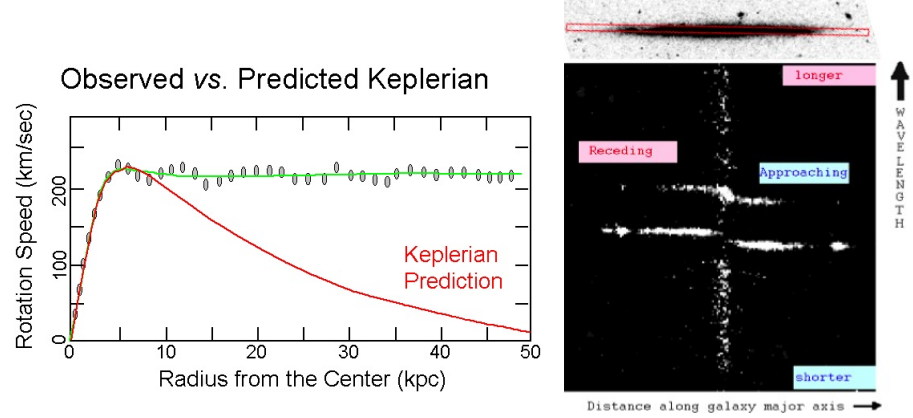


Figure 26.6 *Left:* Data for inferred galactic rotation speed (in km/s, points) vs. radius R from the galactic center (in kpc), with horizontal green line showing data fit that implies a flat, or roughly constant, rotation speed for all radii $R > 5$ kpc. The red curve compares the decline of speed as $V(R) \sim 1/\sqrt{R}$ that is expected from Keplerian motion with a mass that is as centrally concentrated as the stellar light. The difference implies there is a substantial “dark matter” contribution to the mass for $R > 5$ kpc. *Right:* The top panel shows slit exposure for the negative image of an external galaxy viewed with its disk edge-on to the observer line of sight. The lower panel then shows the slit spectrum formed by plotting the wavelength spectrum of the star’s light along the vertical, against distance along the major axis of the galaxy on the horizontal axis. The flat bright emission vs. distance come from Doppler-shifted line emission lines that reflect the galactic rotation away from us on the left (longer wavelength) and toward us on the right (shorter wavelength). The flatness now shows quite directly that the rotation curve of this galaxy is also flat, as in our own Milky Way, again implying the presence of dark matter.

would be likewise contained within this central bulge. But this would then require that galactic orbital speeds should follow the same radial scaling as derived for orbits around other central concentrations of mass, like the planets around the Sun. These follow the standard Keplerian scaling,

$$V_{kep}(R) = \sqrt{\frac{GM}{R}} \sim \frac{1}{\sqrt{R}}, \quad (26.6)$$

which would thus decline with the inverse square root of the radius.

Instead, the constant orbital speed V_o of a flat rotation curve implies that the amount of *mass within a given radius* must *increase* in proportion to the radius³,

$$M(R) = \frac{V_o^2 R}{G} \sim R. \quad (26.7)$$

³ For a spherical distribution of mass, the gravitational acceleration at any given radius R is just set by the mass $M(R)$ *within* that radius. For the visible mass in galactic disk the overall gravitational field is more complex. But in practice, use of the simple spherical scaling form still provides a good approximation for mapping the overall distribution of *dark matter*, which is inferred to have a nearly spherical distribution in the galactic halo.

Since this extra gravitational mass extends to regions with very little luminosity, i.e. that are effectively very dark, it is known as *dark matter*. From studies extending up to scales well beyond our galaxy, to clusters and superclusters of external galaxies, it is now thought that there is about *five* times more dark matter in the universe than the ordinary luminous matter that makes up stars, ISM gas and dust, planets, and indeed us. The origin and exact nature of this dark matter is not known, but it is thought to interact with other matter mainly just through gravity, and not through the electromagnetic and (strong) nuclear force that plays such a key role in the properties of ordinary “baryonic” matter⁴

Nonetheless, as discussed below, this dark matter is now thought to be crucial to the formation of large scale structure in the universe, and thus the associated galaxies that in turn provide the sites for formation of the stars, our Sun, and the planets like our Earth. In short, without dark matter, we wouldn’t be here today to wonder about it!

26.4 Super-massive black hole at the galactic center

The center of our galaxy is in the direction of the constellation Sagittarius, at a distance of about 8 kpc. Over this distance the absorption by gas and dust in the disk plane contribute to some $A_V \approx 25$ magnitudes of visual extinction. This corresponds to a reduction factor $F_{\text{obs}}/F_{\text{int}} \approx 10^{-A_V/2.5} = 10^{-10}$ in the visible flux, so almost completely obscuring this galactic center in the visible parts of the spectrum. But at longer wavelengths in the infra-red and radio, for which the dust opacity is much lower, it becomes possible to see fully into the galactic center. Particularly noteworthy is Sagittarius A* (a.k.a. Sgr A), a region of very bright radio emission.

Quick Question 3: Angular vs. Physical Sizes at the Galactic Center

At the distance $d \approx 8$ kpc of the Galactic Center:

- What is the physical size s (in AU) of an angle $\alpha = 1$ arcsec?
- What is the angular radius α_c (in arcsec) of the central parsec cluster?

Infrared observations of the region around Sgr A shows a concentration of several hundred stars known as the “central parsec cluster”. The blurring effects of the Earth’s atmosphere normally limit spatial resolution to angular sizes of order an arcsec. But using specialized techniques – known as “speckle imaging” and “adaptive optics” – it has become possible over the past couple decades to obtain IR images of individual stars within the central *arcsec* of the cluster, with angular resolution approaching ~ 0.1 arcsec.

Monitoring of the couple dozen stars within this field since the mid-1990’s has revealed them to have small but distinctive *proper motions*, following *curved*

⁴ Ordinary matter is often referred to as “baryonic” because most its mass comes from the protons and neutrons that are generally known as “baryons”. Technically though, a small fraction the mass of ordinary matter comes from electrons, which are actually classified as “leptons”, not a baryons.

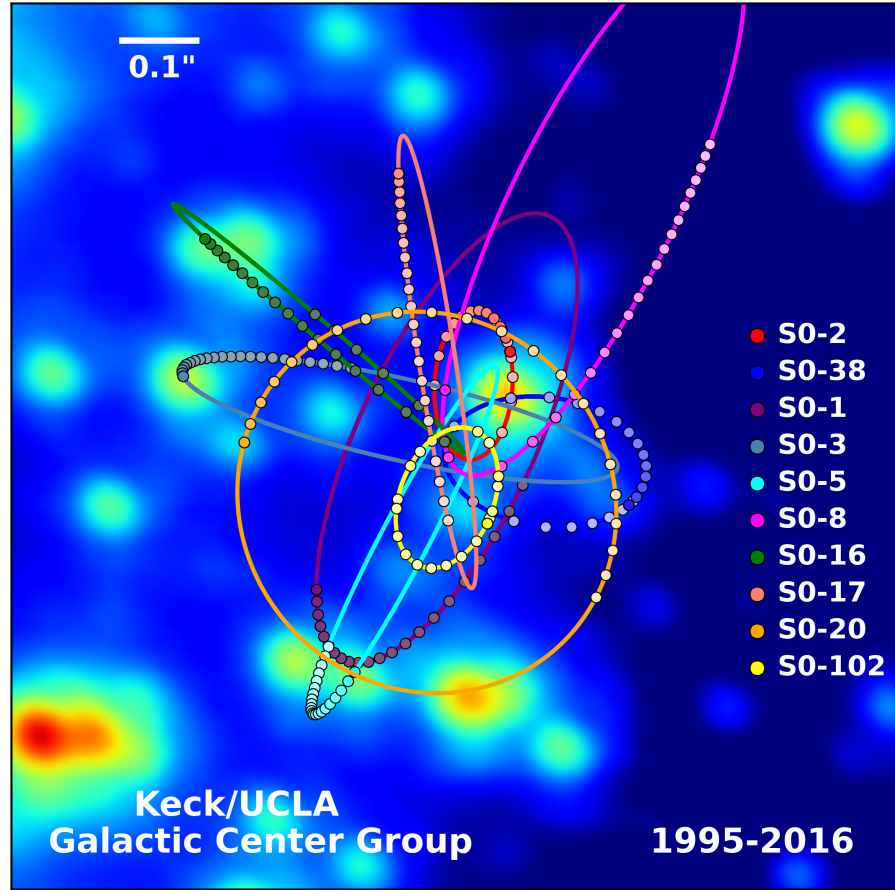


Figure 26.7 Sub-arc resolution of central arcsec of Sgr A, showing orbital tracks of individual stars about a common central point, now identified as the location of a supermassive black hole of mass $M_{bh} \approx 4 \times 10^6 M_{\odot}$. The individual dots show the annual positions of individual stars identified by the color code legend over the 21-year timespan 1995-2016. The star S0-2 has the shortest period, 16.7 years, and so has been tracked over more than a full orbit. This image was created by Prof. Andrea Ghez and her research team at UCLA from data sets obtained with the W. M. Keck Telescopes.

orbital tracks that all center around a common point just slightly offset from the Sgr A radio source. Figure 26.7 illustrates the tracks of 10 stars over the 21-year period 1995-2016, with annual positions of the stars marked by dots, and individual stars identified by the color legend at the right. Using the known $d = 8$ kpc distance, the angular sizes of the orbital tracks can be translated into physical sizes for the semi-major axes a of the orbits. Extrapolating (or following) the motion over a full cycle allows one to infer the orbital periods P . Application of this and the semi-major axis into Kepler's third law then gives an extremely

large mass, $M_{bh} \approx 4 \times 10^6 M_{\odot}$ for the central attracting object, which is inferred to be a *super-massive black hole* (SMBH). Exercise 1 illustrates the process for this mass determination.

Exercise 1: *Using Kepler's 3rd law to infer mass of SMBH*

To compute the mass of the SMBH at the galactic center, consider the star labeled SO-02 in figure 26.7, which has recently completed a full, monitored orbit.

- a. Using the arrow-key in the upper left showing the angular scale, estimate the angular extent (in arcsec) of SO-02's projected major axis.
- b. Using the known distance $d = 8 \text{ kpc}$, what is the associated physical size s (in AU) of the *semi*-major axis of SO-02's orbit?
- c. Next count the number of dots around the orbit to estimate the period P (in yr) of SO-02's orbit.
- d. Assuming we have a face-on view of SO-2's orbit, now use Kepler's 3rd law to estimate the mass M_{bh} (in M_{\odot}) of the central Black hole about which SO-2 is orbiting?
- e. Suppose our view is off by a modest inclination angle i from face-on. Does this increase, decrease, or have no effect on the mass estimate in part d? If it changes, by what factor?

More information on these stars in the central pc can be found at the website for the UCLA Galactic Center Group:

<http://www.galacticcenter.astro.ucla.edu/>

27 External Galaxies



Figure 27.1 The Andromeda galaxy (a.k.a. M31), the nearest large, external galaxy, at a distance of about 2 Mly from our Milky Way.

27.1 Cepheid variables as standard candle for distances to external galaxies

What we now know as external galaxies, like our Milky Way but far outside of it, were first identified by their signature spiral form. Known merely as “spiral nebulae”, it was once thought they might be just stellar or cluster size regions like Planetary Nebulae, or the various other forms of diffuse nebulae seen in association with stars or star clusters.

The situation advanced considerably once telescopes became powerful enough to resolve individual stars within the great spiral nebula in Andromeda. In the 1920's, using the 100-inch telescope on Mt. Wilson, Edwin Hubble was able to observe a particular kind of *pulsating* luminous giant star known as a *Cepheid*

variable. Previous studies of Cepheid variables in our own Galaxy showed that they have the rather peculiar but very useful property that the *period* P of their pulsation – which can be readily measured – is related to their intrinsic *luminosity* L . Using a Cepheid with a measured period as a luminous *standard candle* with a known luminosity, Hubble's observation of the apparent brightness F (actually apparent magnitude m) of Cepheid stars within the Andromeda nebula led him to estimate its distance using the usual standard-candle formula,

$$d = \sqrt{\frac{L}{4\pi F}} = 10^{1+(m-M)/5} \text{ pc}. \quad (27.1)$$

The second equality uses the *distance modulus*, given by the difference between apparent and absolute magnitude of the Cepheid star, with the former observed and the latter inferred from the observed period and the Cepheid Period-Luminosity relation. The results indicated Andromeda was more than a million light years away! Since the Milky Way had already been inferred to have a diameter of only 100,000 light years, it was thus clear that Andromeda must lie well outside our galaxy, indeed with an angular size that implies it has a comparable physical size to the Milky Way itself.

Since this original application of Cepheid variables as standard candles, it has become clear that there are actually two distinct Cepheid classes: Types I and II, which apply respectively to Population I and II stars, with high and low metallicity. Figure 27.2 plots $\log L/L_\odot$ vs. P (days) for Type I and Type II Cepheids, showing that the former are about a factor four more luminous at a given period. Hubble incorrectly assumed that the Cepheids he initially used were of Type II, but they were actually of Type I. Accounting for the factor four higher luminosity within the observed apparent brightness implies that the Andromeda galaxy is actually twice as far as Hubble thought, i.e. some 2 Mly.

27.2 Galactic redshift and Hubble's law for expansion

As Hubble applied his Cepheid method to measuring distances to other spiral nebulae, a Mt. Wilson observatory night assistant named Milton Humason, a former mule driver without even a high-school diploma, became especially skilled at measuring their spectra from very faint images on photographic plates. In particular, he was able to measure the Doppler shift of known spectral lines, giving then a direct measure of the galaxies' radial velocity V_r .

Quite surprisingly, Humason found that, with the exception of the relatively nearby galaxies like Andromeda, all the more distant galaxies showed only *red-shifted* spectral lines, implying from the Doppler shift formula that they are all moving *away* from us, with $V_r > 0$.

Even more remarkably, when combined with Hubble's measurement of galactic distances, it led to what is now known as the *Hubble law*¹ by a *linear* propor-

¹ This is now sometimes referred to the Hubble-Lemaitre Law, because in 1927, two years

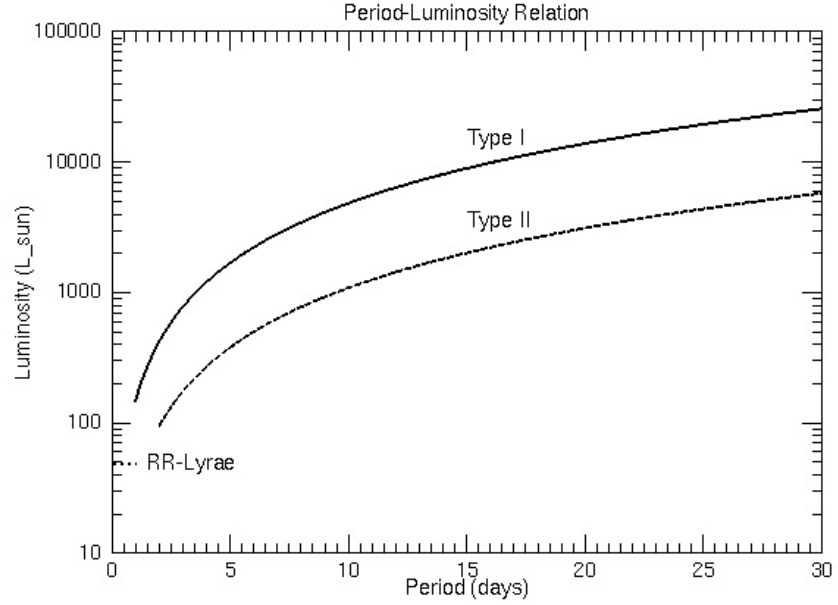


Figure 27.2 Period (in days) vs. Luminosity (in L_{\odot}), on a log scale) for Cepheid variables of Type I (high metallicity, of Population I, upper curve) and Type II (low metallicity, of Population II, lower curve).

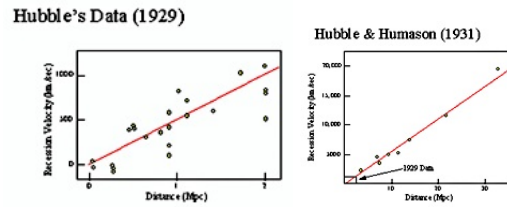


Figure 27.3 The original discovery forms of Hubble's law, showing a roughly linear proportionality between the recession velocity V_r of a galaxy and its distance. The slope of the red line fits through the data points gives a measure of the Hubble constant, $H_o \approx 500$ (km/s)/Mpc. The modern best value is much smaller, $H_o = 67$ (km/s)/Mpc.

tionality between velocity V_r and distance d ,

$$V_r = H_o d. \quad (27.2)$$

The proportionality constant, H_o , is known as the *Hubble constant*, which has units of an *inverse time*. Figure 27.3 plots the original relations obtained by Hub-

before Hubble published his own article, the Belgian priest and astronomer Georges Lemaitre published an article describing the law, but in French in a relatively obscure journal. It was thus overlooked until translated into English in 1931, two years after Hubble's paper.

ble and Humason, with the *slope* of the red line fit to the data points giving H_o . Because of the incorrect assumption of the Cepheid type, along with a combination of other errors, the original value of nearly $H_o \approx 500$ (km/s)/Mpc turns out to be a serious overestimate of the modern best value of $H_o \approx 70$ (km/s)/Mpc.

The implications of this Hubble law are truly profound. In particular, if we simply assume that the velocity is constant, then dividing the distance by velocity gives the time since a distant galaxy was at zero distance from us,

$$t = \frac{d}{V_r} = \frac{1}{H_o} \equiv t_H \approx 10 \text{ Gyr} \frac{100 \text{ (km/s)/Mpc}}{H_o}, \quad (27.3)$$

where the second equality shows that this time, which is *same* for *all galaxies*, is given by the inverse of the Hubble constant. This is known as the *Hubble time*, $t_H \equiv 1/H_o$, and as shown in the last relation of (27.3), a Hubble constant of $H_o = 100$ (km/s)/Mpc gives a Hubble time of approximately $t_H \approx 10$ Gyr.

Modern observations of very distant galaxies show that the redshift,

$$z \equiv \frac{\lambda_{obs}}{\lambda} - 1 = \frac{V_r}{c}, \quad (27.4)$$

can become quite large, with even some cases having $z > 1$. If taken literally in terms of the latter velocity Doppler shift formula in (27.4), this would seem to suggest that $V_r > c$, in apparent contradiction of special relativity.

But as discussed in the cosmology sections in part 5, a more proper interpretation of this “cosmological redshift” is that it represents the *stretching* of the wavelength of light by the *expansion of space itself*! This can readily lead to redshifts $z > 1$. Einstein’s limit really applies to how fast objects can travel relative to space, but that space itself can expand at a speed *faster* than light!

27.3 Tully-Fisher Relation: $L_{gal} \propto V_{rot}^4$

For more distant galaxies, it becomes increasingly difficult to detect and resolve even giant stars like Cepheid variables as individual objects, limiting their utility in testing the Hubble law to relatively modest distances and redshifts. For much larger distances, we need another, brighter “standard candle”, like the white-dwarf supernovae (WD-SN) discussed in §31.1. But because the unpredictability of their appearance long limited the number of such WD-SN² detections, an important alternative method has been the so-called *Tully-Fisher* relation. Empirically, it was found that the luminosity of a spiral galaxy, L_{gal} , scales with maximum rotation velocity V_{rot} inferred from Doppler shift of spectral lines, with the approximate form

$$L_{gal} \propto V_{rot}^4. \quad (27.5)$$

The proportionality constant depends on the spectral band, but as an example, in the near-infrared “I-band” (centered at 820 nm), the relation in terms of the

² In the standard, formal notation, WD-SN are classified as “Type Ia”, or “SN Ia”

absolute magnitude takes the numerical form,

$$M_I \approx -3.3 - 8.3 \log \left(\frac{V_{rot}}{\text{km/s}} \right). \quad (27.6)$$

Since magnitude $M \propto -2.5 \log L$, the slope of -8.3 here implies a velocity exponent $8.3/2.5 \approx 3.3$, somewhat shallower than the power 4 assumed in eqn. (27.5). Figure 27.4 shows actual I-band magnitude data vs. $\log(V_{rot})$, compared with linear relations with slope -8.3 (blue) and -10 (red), the latter corresponding the standard Tully-Fisher law (27.5) with velocity exponent 4 = 10/2.5.

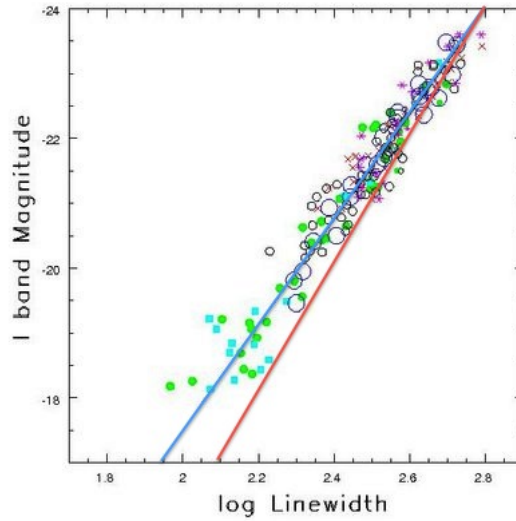


Figure 27.4 Empirical Tully-Fisher relation in the infrared, plotted as I-band absolute magnitude M_I vs. logarithm of the total line-width (in km/s), set by twice the rotational velocity, $2V_{rot}$. The blue line shows best fit line with slope -8.3 , as given in eqn. (27.6), corresponding to velocity exponent $8.3/2.5 \approx 3.3$, slightly shallower than the slope 4 assumed in the standard $\log L_{gal}$ vs. $\log V_{rot}$ scaling law of eqn. (27.5). The red line shows the slope $-10 = -2.5 \times 4$ that would be implied in the I-band magnitude scaling by this standard form for the Tully-Fisher relation. The best-fit slope can differ for different wavebands.

Exercise 1: *Application of Tully-Fisher relation.*

A spiral galaxy with redshift $z = 0.23$ and apparent I-band magnitude $m_I = +17.6$ has an observed total spectral line width ratio, $\Delta\lambda/\lambda = 0.0013$. Compute the galaxy's:

- Orbital velocity V_{rot} ;
- Absolute I-band magnitude M_I ;
- I-band luminosity L_I , in units of the I-band luminosity of the Sun, $L_{I,\odot}$ (for which the I-band absolute magnitude is $M_I \approx +4$).
- Distance modulus $m_I - M_I$;
- Distance D (in Mpc).
- Recession velocity V_r from redshift z ;

g. Associated Hubble constant $H_o = V_r/D$.

To glean a possible physical rationale for this empirical Tully-Fisher relation, first note again that by Kepler's law the rotational velocity V_{rot} at an outer radius R scales with galactic mass M_{gal} as

$$V_{rot}^2 = \frac{GM_{gal}}{R}. \quad (27.7)$$

On the other hand, the galactic luminosity L_{gal} scales with the galaxy surface brightness I_o times the surface area πR^2 out to this outer radius,

$$L_{gal} \propto I_o \pi R^2. \quad (27.8)$$

Combining (27.7) and (27.8) gives the scaling

$$L_{gal} \propto \frac{V_{rot}^4}{I_o (M_{gal}/L_{gal})^2}, \quad (27.9)$$

which recovers the standard Tully-Fisher scaling of eqn. (27.5) *if* we assume a constant value for the surface brightness times the square of the mass-to-light ratio, $I_o (M_{gal}/L_{gal})^2$. Models of galaxy formation have tried to explain why this should be true, but the results are tentative and not clearly established and accepted. Nonetheless, as a strictly *empirically* calibrated relation, this Tully-Fisher scaling provides a luminous standard candle to infer distances beyond the range accessible to the Cepheid method, and so allows a calibration of the Hubble law to moderately large distances and redshifts.

27.4 Spiral, Elliptical, & Irregular galaxies

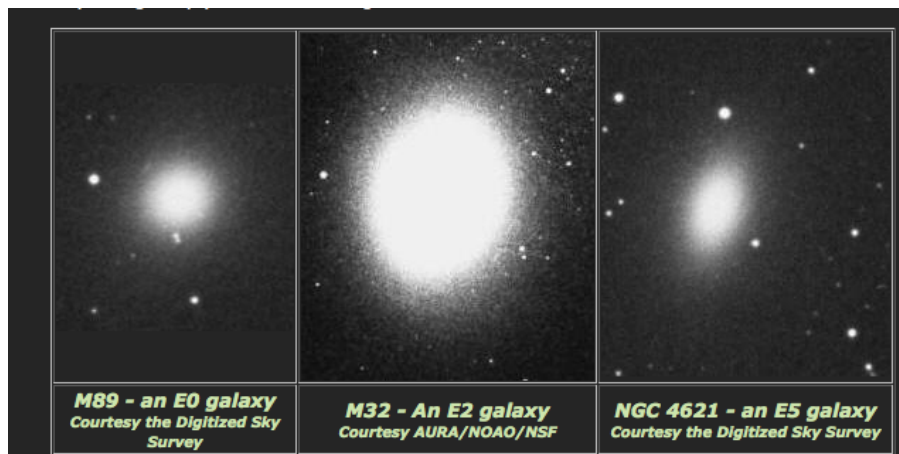


Figure 27.5 Examples of 3 types of elliptical galaxies; figure taken from <http://cas.sdss.org/dr6/en/proj/basic/galaxies/ellipticals.asp>.

Galaxies can be generally classified by three distinct types of morphology: Spiral, Elliptical, and Irregular.

Spiral galaxies are similar to our Milky Way, with distinct disk, halo, and bulge components. A spiral density wave in the disk forms the spiral arms that are the regions of active star formation out of the cold clouds of gas and dust. The tightness of the winding of the arms can vary, and sometimes emanate from a central “bar”. M51, a.k.a. the “Whirlpool” galaxy, shown in figure 21.5, provides a good example of a typical spiral galaxy. As illustrated in fig. 26.2, our Milky Way galaxy is thought to be a barred spiral.

Elliptical galaxies have a spheroidal shape, with different gradations of elongation from nearly spherical (E0) to highly extended (E5), as illustrated in figure 27.5. Their stars are generally found to be Population II, and thus quite old with reduced metallicity. There appears to be a near absence of ISM gas or dust, and thus little or no new star formation. In these respects, elliptical galaxies are similar to globular clusters that orbit in the halo of our Milky way, but much bigger and more massive. Their physical sizes can span a large range, from about 0.1 to 10 times size of the 100,000 ly diameter of our Milky way, i.e. only 10^4 ly for “Dwarf ellipticals” (with $M \sim 10^9 M_\odot$), to 10^6 ly for giant ellipticals (with $M \sim 10^{12} M_\odot$). At the center of a very large cluster of galaxies, there is often a giant, “central dominant” (CD) elliptical galaxy that can have mass of $10^{12} M_\odot$ or more.

Irregular galaxies are just that. The overall structure is complex, though within subareas there can be spiral features. In many cases, it seems likely that the irregular form is because we are actually viewing two colliding galaxies, with then their mutual tidal interaction warping and disrupting whatever symmetric forms may have existed in the source galaxies. Figure 27.6 shows a mosaic of interacting galaxies (right), and a close-up the direct collision underway in the Antenna galaxy (left).

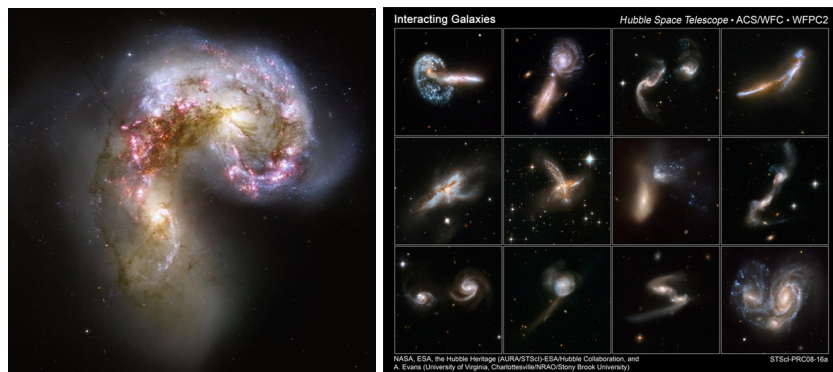


Figure 27.6 *Left:* The antenna galaxy, which is actually two galaxies undergoing a direct collision. *Right:* Gallery of other examples of interacting galaxies.

27.5 Role of Galaxy Collisions

For any collection of objects of size s separated by a mean distance d , the number density is $n \approx 1/d^3$ while the cross section is $\sigma \approx s^2$. The mean-free-path for collision is then

$$\ell = \frac{1}{n\sigma} \approx d \left(\frac{d}{s} \right)^2. \quad (27.10)$$

For individual stars, the distance/size ratio is enormous, of order $d/s \approx \text{pc}/R_\odot \sim 4 \times 10^7$, implying a mean-free-path $\ell \sim 10^{15} \text{ pc}$! Since this is more than 10^{10} larger than the $\sim 30 \text{ kpc}$ size of a galaxy, we can conclude that individual field stars in galaxy should *never* collide³.

But for galaxies, this ratio of distance/size is much smaller, about a factor 20 for us to the Andromeda galaxy, and often just a factor few for galaxy clusters. Moreover, in the early universe, the average separation among galaxies not in the same cluster was smaller, with the factor reduction just set by the redshift, $z + 1$. As such, while somewhat rare in the current-day universe, collisions can and do occur, and they were much more common in the early universe. Indeed, some models invoke a “bottom up” scenario in which larger galaxies form from the merger of smaller galaxies.

Animations from computer simulations of two colliding galaxies can be found on web at:

<http://www.youtube.com/user/galaxydynamics>

The video dubbed “Spiral Galaxy” shows how spiral density waves can be induced by orbiting clumps of dark matter.

When galaxies do collide, their overall pattern of stars become strongly distorted by the mutual tidal interaction of the overall mass of the two galaxies; but the individual stars are too widely separate to collide, and so just pass by each other. In contrast, any gas clouds in the ISM of each galaxy do collide, with the resulting compression increasing the density of gas and dust, and thus often triggering a strong burst of new star formation. Such colliding systems are indeed often dubbed “starburst galaxies”.

While distant galaxies show a redshift that implies they are moving away from us as part of the expansion of the universe, the mutual gravitational attraction between our Milky Way and the relatively nearby Andromeda galaxy is actually pulling them toward each other. Indeed, it now seems likely that Andromeda and the Milky Way will collide in about 3-4 Gyr. The “Future Sky” animation in the above link shows how the sky might appear from Earth during this collision.

³ Some gravitational interaction can occur in the dense cores of compact globular clusters, but even there direct collision between stars is very unlikely.

28 Active Galactic Nuclei (AGNs) and Quasars

28.1 Basic properties and model

During the 1960's sky surveys with radio telescopes discovered "QUAsi-Stellar Radio" sources, now known as "Quasars", or also Quasi-Stellar Objects (QSOs). In contrast to the extended radio emission sources seen from various regions of the galaxy, these QSOs are, like stars, *point*-like sources without any readily discernible angular extent. They were soon identified with similarly point-like sources in the visible and other wavebands. But quite *unlike* stars, their spectral energy distribution does *not* even roughly match that of a Black-body of any temperature; instead it has an extended power-law form over a wide range of energies from the radio through the IR, visible, UV and even extending to the X-ray and gamma-rays.

Nonetheless, this broad spectral distribution does still show patterns of absorption (and emission) lines that can be identified with known elements, but notably with a *huge redshift* z . For example, 3C273, one of the first and most famous QSOs, has $z = 0.158$, meaning that it is receding from us at a radial speed $v_r = 0.158c \approx 47,000$ km/s. Taking a Hubble constant $H_o \approx 70$ (km/s)/Mpc, this puts its distance at $d \approx v_r/H_o \approx 677$ Mpc ≈ 2.1 Gly. With associated distance modulus $m - M = 5 \log(d/10 \text{ pc}) \approx +39$, together with its apparent magnitude $m = +15$, this implies an absolute magnitude $M \approx -24$, or luminosity $L \approx 5 \times 10^{11} L_\odot$! This far exceeds the luminosity of any star, and indeed even outshines the luminosity of a typical galaxy of $\sim 10^{11} L_\odot$.

Modern observations, e.g. with the Hubble Space Telescope, revealed that these quasars are commonly surrounded by a comparatively faint, diffuse stellar emission from a host galaxy. It is now realized that QSOs are indeed just one example of a class of "Active Galactic Nuclei" (AGNs). In contrast to the extended galactic emission over a distance of a galactic diameter ~ 30 kpc, QSO/AGN emission is entirely point-like, emanating from the galactic nucleus. Indeed, since such QSO/AGNs often vary over time scales as short as a day, they must be very compact, no more than a light-day in diameter, or $\lesssim 100$ AU; this means they are roughly of order $\sim 10^8$ (~ 30 kpc/100 AU) times smaller than their host galaxy.

This extreme luminosity from such a small volume is thought to be the result of matter accreting onto the supermassive black hole (SMBH) at the center of the QSO/AGN host galaxy. The SMBH in our Milky Way, and indeed in

most galaxies in the nearby, current-day universe, are relatively inactive, with relatively little ongoing accretion. But in the early universe, when the smaller inter-galactic separation meant more frequent galaxy collisions, the extreme disruption caused some stars to approach so close to the SMBH that they became tidally disrupted. The remnant stellar material typically still had too much angular momentum to fall directly onto the SMBH, and so instead fed an *accretion disk*. The viscous shear transports angular momentum outward, allowing a steady, gradual accretion in which the gravitational energy released heats the disk and powers its emitted luminosity.

Quick Question 1: *Energy efficiency for accretion near a black hole*

For accretion down to a radius that is a factor R_{acc}/R_s times the Schwarzschild radius of black hole, compute the energy efficiency factor $\epsilon = E_g/mc^2$ for the gravitational energy gain E_b as a fraction of the rest mass energy mc^2 of the accreted mass. Confirm that $\epsilon = 0.1$ for $R_{acc}/R_s = 5$.

Accretion down to the vicinity of a black hole can generate energy that is a substantial fraction of the rest mass energy of the accreting matter. (See QQ 1.) For an accretion rate \dot{M}_{acc} and conversion efficiency ϵ , the generated luminosity is

$$L_{acc} = \epsilon \dot{M}_{acc} c^2 = 1.4 \times 10^{12} L_{\odot} \frac{\epsilon}{0.1} \frac{\dot{M}_{acc}}{M_{\odot}/yr}. \quad (28.1)$$

The second equality shows the very enormous luminosity associated with accretion of $1 M_{\odot}/yr$ at a efficiency of $\epsilon = 0.1$ (the value for accretion to 5 Schwarzschild radii). It indeed readily equals or exceeds the luminosity inferred from the observed apparent magnitude and estimated distance of QSOs, including the example of 3C 273 mentioned above.

The SMBHs that power quasars are thought to be even more massive than those found in our and other nearby galaxies, of order a *billion* solar masses ($10^9 M_{\odot}$). But the associated Schwarzschild radii, $R_s \sim 3 \times 10^9 \text{ km} \sim 20 \text{ AU}$ are still small enough to accommodate the day-timescale variation, even accounting for the fact that the emission region is likely to extend over $5 - 10 R_s$.

28.2 Lyman alpha clouds

As this enormous luminosity from distant quasars propagates through the universe, it can sometimes pass through the relatively higher-density gas associated with galaxies or a galaxy cluster. Since the quasar spectral distribution extends well into the UV, the photons at wavelengths $\lambda = 121.57 \text{ nm}$ for the Lyman alpha (Ly- α ; $n = 1$ to $n = 2$) transition of neutral Hydrogen, for which the opacity is very high, become strongly absorbed. But along this extended path length, the local Ly- α wavelength is Doppler shifted by the Hubble expansion, extending it to a longer wavelength that depends on the distance to the absorbing intergalactic H-cloud. As illustrated in figure 28.1, this makes the observed quasar

spectrum have a distinct number of absorption lines. Indeed, sometimes these are so dense that they are known as the *Lyman-alpha* “forest”, with each “tree” of the forest corresponding to a distinct inter-galactic H cloud at a distance set by the cosmological (Hubble-law) red-shift of that observed absorption feature.

In essence, the huge luminosities and huge distances of quasars provide us a set of “flashlights” to probe the inter-galactic Hydrogen gas in the universe between us and the quasars.

Quick Question 2: *Lyman cloud speed and distance*

Suppose a quasar shows absorption from a Lyman-alpha cloud at an observed wavelength $\lambda_{obs} = 183 \text{ nm}$.

- What is the redshift z for this cloud.
- What is its inferred recession speed v_r ?
- For a Hubble constant $H_o = 67 \text{ (km/s)/Mpc}$, what is its distance?

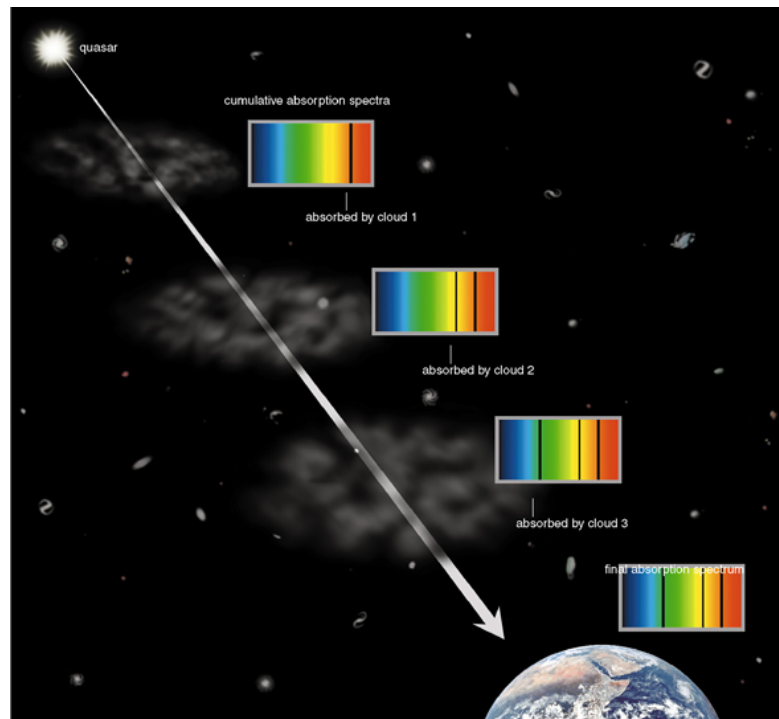


Figure 28.1 Illustration of Lyman- α clouds, in which Hydrogen gas in galaxies at various redshifts absorb distant quasar light in the Lyman- α line from the $n = 1$ to $n = 2$ transition of Hydrogen.

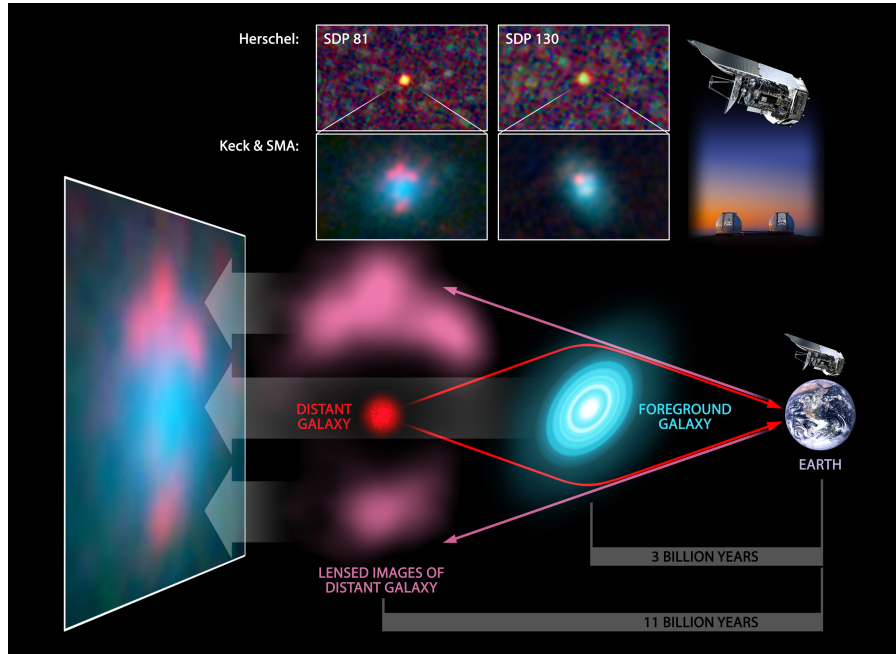


Figure 28.2 Diagram to illustrate “gravitational lensing”, wherein the image of a distant galaxy or quasar is multiplied by the gravitational bending of light from passage near the very large mass of an intervening galaxy or galaxy cluster.

28.3 Gravitational lensing of quasar light by foreground Galaxy Clusters

As this enormous luminosity from distant quasars propagates through the universe, it can also sometimes pass so close to a galaxy cluster that the gravity from the cluster’s mass actually *bends* the rays of light, forming what is known as a “gravitational lense”. This basic effect of gravitational bending of light was predicted by Einstein’s General Theory of Relativity, and was famously confirmed by expeditions to measure the associated shift in the position of stars as their light passed near the Sun during a solar eclipse. In the context of the passage of quasar light by a galaxy cluster, it can lead to multiple images, or even an “Einstein arc” or circle, if the quasar and galaxy’s mass-center are both closely aligned with the observer’s light of sight. Figure 28.2 illustrates the basic geometry and process.

From General Relativity, the bending angle θ depends on the mass M of the lens and the “impact distance” b of the light ray from the background source passing the lensing mass,

$$\theta = \frac{4GM}{bc^2} . \quad (28.2)$$

The factor 4 is a general relativistic correction factor for the simple Newtonian

calculation for bending of an object with incoming speed $V_x = c$, as illustrated in figure 28.3.

Gravitational Bending of Light

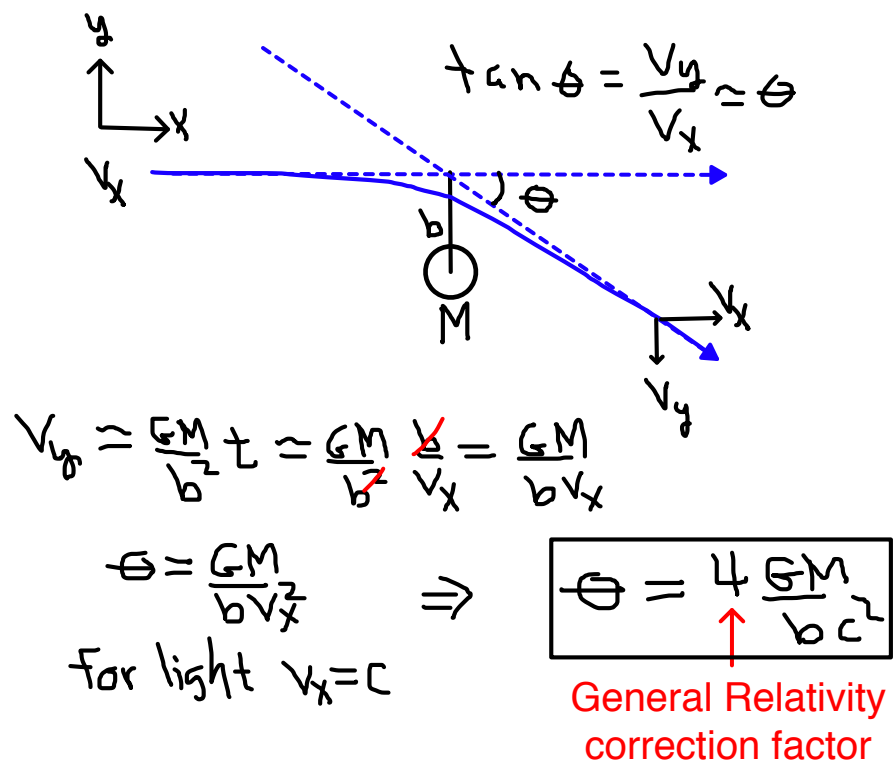


Figure 28.3 Diagram to illustrate the gravitational deflection of an object with initial speed V_x impacting within a distance b of a gravitational mass M . The same scaling applies to light with speed $V_x = c$, but with a correction factor 4 derived from General Relativity, giving then the correct scaling for gravitational bending of light.

Exercise 1: Gravitational Lensing

Suppose a distant galaxy cluster with redshift $z = 0.2$ has two identical quasar images at equal angles $\theta = 10$ arcsec on each side of the cluster center.

- Assuming the current best value for Hubble constant, $H_0 = 67$ (km/s)/Mpc, what is the distance D (in Mpc) to the lensing galaxy?
- Use this distance D and the angle θ to estimate the closest distance b (in kpc) that the quasar's light passes to the center of the galaxy.
- Now use this b and the angle θ in Einstein's gravitational lensing formula to estimate the mass M (in M_\odot) of the lensing galaxy cluster.

28.4 Gravitational redshift

Another effect related to gravitational lensing is the *gravitational redshift* experienced by light emitted from a radius R_o near a gravitational mass M , which from General Relativity is given by

$$z = \frac{1}{\sqrt{1 - R_s/R_o}} - 1, \quad (28.3)$$

where $R_s \equiv 2GM/c^2$ is the Schwarzschild radius for the mass M . (See §18.2.4 and eqn. 18.10.) As illustrated in figure 28.4, for the non-relativistic case that the initial radius is far above the Schwarzschild radius, $R_o \gg R_s$, straightforward Taylor expansion leads to a simple form that casts this photon redshift in terms of simple conservation of total energy of the photon plus gravity, with the loose association of an equivalent initial photon “mass” $m_o = E_o/c^2$ based on Einstein’s energy-mass equivalence principle.

Exercise 2: Gravitational redshift as alternative explanation of quasar redshift

Suppose we try to explain the redshift of 3C273 ($z=0.158$) as a gravitational redshift, rather than being from cosmological expansion.

- Relative to Schwarzschild radius R_s , from what radius R_o is the radiation emitted?
- If the width of lines is 0.1% of their central wavelength, what is the range of radii (relative to R_s) from which the radiation can be emitted.
- For a more physically reasonable assumption that any emission would come from at least radius range $\pm 10\%$ around the central radius R_o , what would be the relative width $\Delta\lambda/\lambda_o$ of the observed emission line?

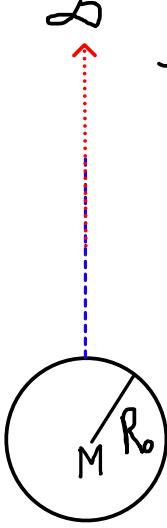
28.5 Apparent “super-luminal” motion of quasar jets

The accretion that powers the tremendous luminosity of quasars can also drive relativistic jets from the polar axes perpendicular to the accretion disk. The jets emit in energies from the radio to gamma-rays, but can be most finely resolved (to angular resolution less than a *milli-arcsecond*!) spatially in the radio, using *Very Long Baseline Interferometry* (VLBI) from multiple radio telescopes spread across the Earth. Indeed, such VLBI radio observations of such jets show they can be quite clumpy and variable on time scales of weeks to years. Quite remarkably, individual clumps in these quasar jets can sometimes show an apparent “super-luminal” motion, meaning that, for the inferred quasar distance, the propagation of individual jet clumps away from the quasar can *appear* to be *faster* than the speed of light!

As illustrated in figure 28.5, the motion is actually a fraction $\beta = v/c \lesssim 1$ that is near but below light speed c , but with a direction toward the observer that changes the light travel time in such a way to make it *appear* the transverse motion is faster than light. For *actual* light speed fraction $\beta < 1$ at an angle θ

Gravitational Redshift of Light

from General Relativity:



$$z+1 = \frac{\lambda_\infty}{\lambda_0} = \frac{\nu_0}{\nu_\infty} = \frac{E_0}{E_\infty} = \frac{1}{\sqrt{1 - \frac{R_s}{R_0}}}$$

$$\text{For } R_s = \frac{2GM}{c^2} \ll R_0$$

$$\sqrt{1 - \frac{R_s}{R_0}} \approx 1 - \frac{R_s}{2R_0} \approx 1 - \frac{GM}{c^2 R_0} = \frac{E_\infty}{E_0}$$

$$E_\infty = E_0 - \frac{GMm_0}{R_0} \quad m_0 = \frac{E_0}{c^2}$$

Figure 28.4 Diagram to illustrate the gravitational redshift of light emitting from an initial radius R_0 near a mass M . By general relativity, the redshift depends on the ratio of the Schwarzschild radius R_s to the initial radius R_0 , but for non-relativistic cases with $R_s \ll R_0$, this just reduces to a simple conservation of total energy of the photon as it climbs out of the gravitational potential of the mass M . The small gravitational redshift from light emitted upward in terrestrial laboratories has been extensively confirmed using laser experiments.

from the direction to the observers, the *apparent* light speed fraction is given by

$$\beta_{app} = \frac{\beta \sin \theta}{1 - \beta \cos \theta}. \quad (28.4)$$

The special case with $\beta = \cos \theta$ gives the maximum apparent speed, which in units of the speed of light is

$$\beta_{app}^{max} = \frac{\beta}{\sqrt{1 - \beta^2}}. \quad (28.5)$$

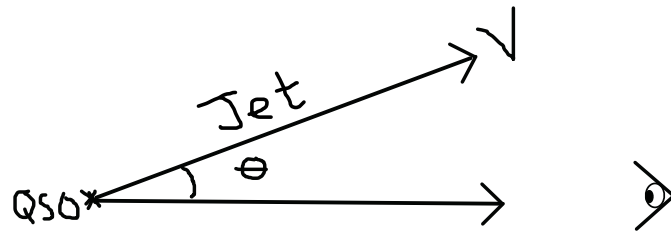
From this it is clear that apparent super-luminal propagation $\beta_{app}^{max} > 1$ is possible whenever the propagation speed $v = c\beta > c/\sqrt{2} = 0.707c$.

Exercise 3: Apparent super-luminal motion in a quasar

Suppose VLBI radio monitoring shows that over 5 years a jet subcomponent of a quasar at a known distance $D = 1$ Gpc has moved away from the center by an angle $\Delta\alpha = 10^{-3}$ arcsec.

-
- a. For the quasar distance D , what is the inferred apparent transverse speed of this component compared to the speed of light, $\beta_{app} = V_{app}/c$?
 - b. What is the minimum actual fraction of the speed of light $\beta = V/c$ needed to give this apparent super-luminal speed β_{app} ?
 - c. What is the associated angle θ (in radians and degrees) between the component’s motion and our line of sight?

Apparent Super-Luminal motion of QSO jets



$$\beta_{app} = \frac{V_{app}}{c} = \frac{\cancel{v} \sin \theta / c}{\cancel{1} - \frac{\cancel{v}}{c} \cos \theta} = \boxed{\frac{\beta \sin \theta}{1 - \beta \cos \theta} = \beta_{app}}$$

$$0 = \frac{d\beta_{app}}{d\theta} = \frac{\cancel{\beta} \cos \theta - \beta \sin^2 \theta}{\cancel{1 - \beta \cos \theta} (1 - \beta \cos \theta)^2} = 0$$

$$0 = \cancel{\cos \theta} - \cancel{\beta} \cos \theta - \beta + \cancel{\beta} \cos^2 \theta \Rightarrow \text{max at } \boxed{\cos \theta = \beta}$$

$$\beta_{app}^{max} = \frac{\beta \sqrt{1 - \beta^2}}{1 - \beta^2} = \boxed{\frac{\beta}{\sqrt{1 - \beta^2}} = \beta_{app}^{max}}$$

Figure 28.5 Derivation to show how quasar jet motion near the speed of light in a direction tilted toward the observer (on right) can lead to an apparent “super-luminal” (faster than light) propagation speed away from the quasar. The maximum apparent speed β_{app}^{max} (in units of the light speed c) occurs when the actual light speed fraction $\beta = v/c < 1$ equals $\cos \theta$, the projection of the jet direction toward the observer.

29 Large Scale Structure and Eras in the Evolution of the Universe

29.1 Galaxy clusters & super-clusters

Much as stars within galaxies tend to form within stellar clusters, the galaxies in the universe also tend to collect in groups, clusters, or even in a greater hierarchy of clusters of clusters, known as “*super-clusters*”. Our own Milky Way is part of a small cluster known as the “Local Group”, which includes also the Andromeda galaxy, as well as up to several dozen smaller, “dwarf” galaxies. Along with roughly a hundred or so other groups, this makes up the “Local Supercluster”, with the highest concentration in the direction of the constellation Virgo. That concentration is also known as the Virgo (super)cluster, at a distance of about 20 Mpc, but it’s outer extent could be even be defined to include the Local Group, making it a possible center of the local super-cluster. In any case, this is just one of millions of super-clusters in the known universe.

Over the past couple decades there have been several very large surveys that aim to measure the redshift of a large number (nowadays reaching many *millions*!) of galaxies along selected swaths of the sky. Over these large expanses of the universe, this measured redshift z gives, for a known Hubble constant H_o , a direct measure of the distance D to the galaxy,

$$D \approx z \frac{c}{H_o} = z D_o \quad ; \quad z \ll 1, \quad (29.1)$$

where the latter equality defines the “*Hubble distance*” $D_o \equiv c/H_o$, which is just the distance that light travels over a characteristic Hubble time, $t_H \equiv 1/H_o$ (cf. §27.2). (This simple relation only applies for modest redshifts $z \ll 1$; as discussed in part V on Cosmology, for $z > 1$, there is a more general relation in terms of the change in the universe’s scale factor $R(t)$.)

With the readily measured two-dimensional (galactic longitude and latitude) positions on the sky, a 2D survey along a swath on the sky can be combined with the redshift distance to form a 3D picture of the universe through that swath. The upper left panels of figure 29.1 (blue color) show a slice of this 3D picture containing one dimension of galactic position plus the distance, arranged along the radius from our own observer’s position at the origin. The result shows a remarkable “cosmic web” in the overall *large-scale structure* (LSS) of the universe. This has a concentration of galaxies along extended, thin “walls”, surrounding huge *voids* with few or no galaxies in the huge volume *between* the walls. But

there are particularly high concentrations at the *intersections* of the walls. Indeed, most previously identified super-clusters can be associated with one of these wall intersections.

The lower right panels of figure 29.1 (red color) show the results of very large simulations for the formation of the structure from the gravitational attraction by matter. For one such simulation, figure 29.2 shows a sequence of volume renderings at different stages of the formation of structure, identified by the redshift z associated with each epoch. As shown in the upper left panel for the earliest phase of the simulations at a redshift $z = 27.30$, one also requires a small initial seed of density fluctuations, which are then amplified by the gravitational attraction. As discussed in part V on cosmology, it is now thought that this initial seed of small-amplitude variations in density is provided by quantum fluctuations in the very early phases of the big-bang itself!

29.2 Dark matter: Hot vs. Cold, WIMPs vs. MACHOs

A key result of these simulations is that achieving an LSS that has the same statistical form as the observed structure requires inclusion of a significant component of *cold dark matter* (CDM), with a total mass that is factor several (~ 5) times the mass of ordinary matter that makes up planets, stars, and galaxies that produce the various spectral bands of electromagnetic radiation that we can directly observe. “Cold” here means that the matter is non-relativistic, so that its gravitational contribution comes from its rest mass, and not from any relativistic enhancement in its energy. Only CDM seems able to form the deep gravitational wells from mutual attraction to frame the observed wall+void network of galaxies observed for large-scale-structure. Hot dark matter tends to remain too distributed.

There are two candidates for CDM, dubbed by the somewhat whimsical terms “WIMPs” – for Weakly Interacting Massive Particles – and “MACHOs” – for Massive Compact Halo Objects. The latter refer to a conjectured large population of low-mass objects – perhaps roving Jupiter size bodies that are too cold to emit much radiation – thought to occupy the core and halo of galaxies. To explain the flat rotation curves of galaxies, the number density of such MACHOs would have to be so large that as they randomly pass in front of stars they should induce a gravitational “micro-lensing” event that should be observable from monitoring of the star’s light. Extensive monitoring surveys have indeed detected such micro-lensing events, but at a rate that is well below what would be needed for MACHOs to be a significant component of dark matter mass.

There is thus now a general consensus that CDM most likely consists of some kind of WIMP. “Weakly interacting” in this context means they are *not* subject to either the *strong nuclear* force, which binds the nucleus of atoms, or *electromagnetic* forces, which bind electrons to atoms, and are responsible for producing light and all other forms of electromagnetic radiation. The inability to produce

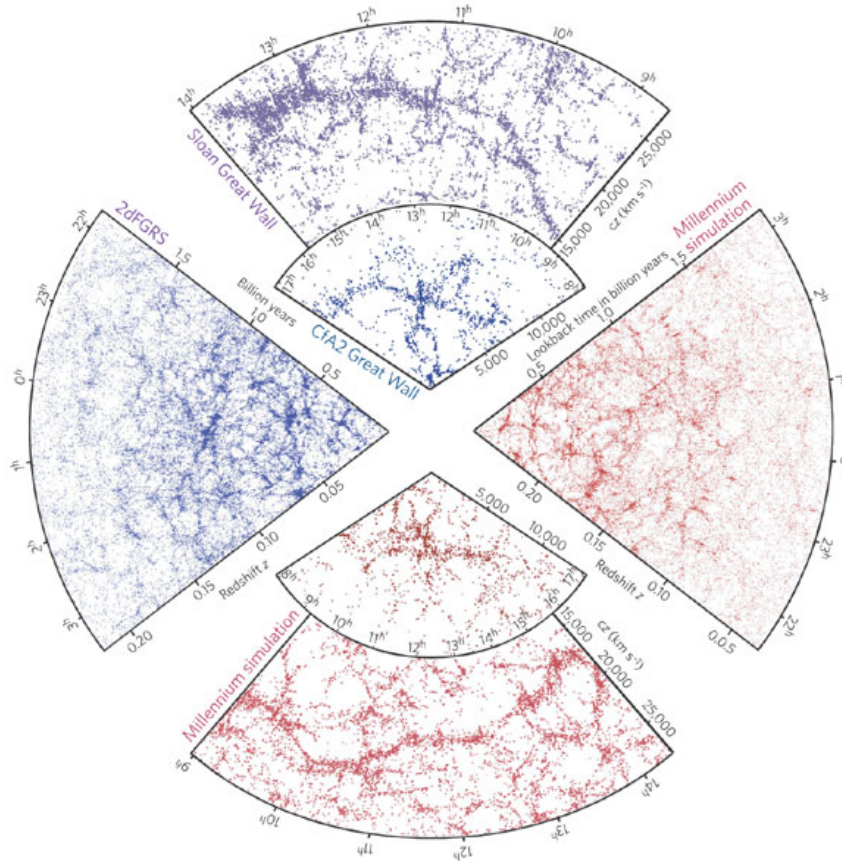


Figure 29.1 Comparison between observational surveys (blue, top and left) vs. computer simulations (red, bottom and right) of the large scale structure of the local universe. The observational surveys measure position and redshifts of millions of galaxies along an extended, narrow arc of the sky, using the redshift z to estimate galactic distance. The simulations assume initial seed perturbations set to correspond to those inferred from Cosmic Microwave Background (CMB) fluctuations, plus cold dark matter to enhance gravitational attraction, and then compute gravitational contraction of structure starting from nearly uniform early universe at redshift $z > 25$ to the present, highly structured, local universe with redshift $z < 0.25$.

light is indeed what makes WIMPs a candidate for dark matter. Like neutrinos, they might be subject to the *weak* nuclear force, but otherwise they only interact with ordinary matter via gravity. Moreover, while neutrinos have a rest mass only a few eV, a hypothetical WIMP could have a much larger mass, perhaps many hundreds times the GeV mass of protons and neutrons. There are several projects underway to detect WIMPs, through experiments deep underground, which shields against the flux of cosmic rays that would otherwise contaminate detections of the very few weak interactions by WIMPs.

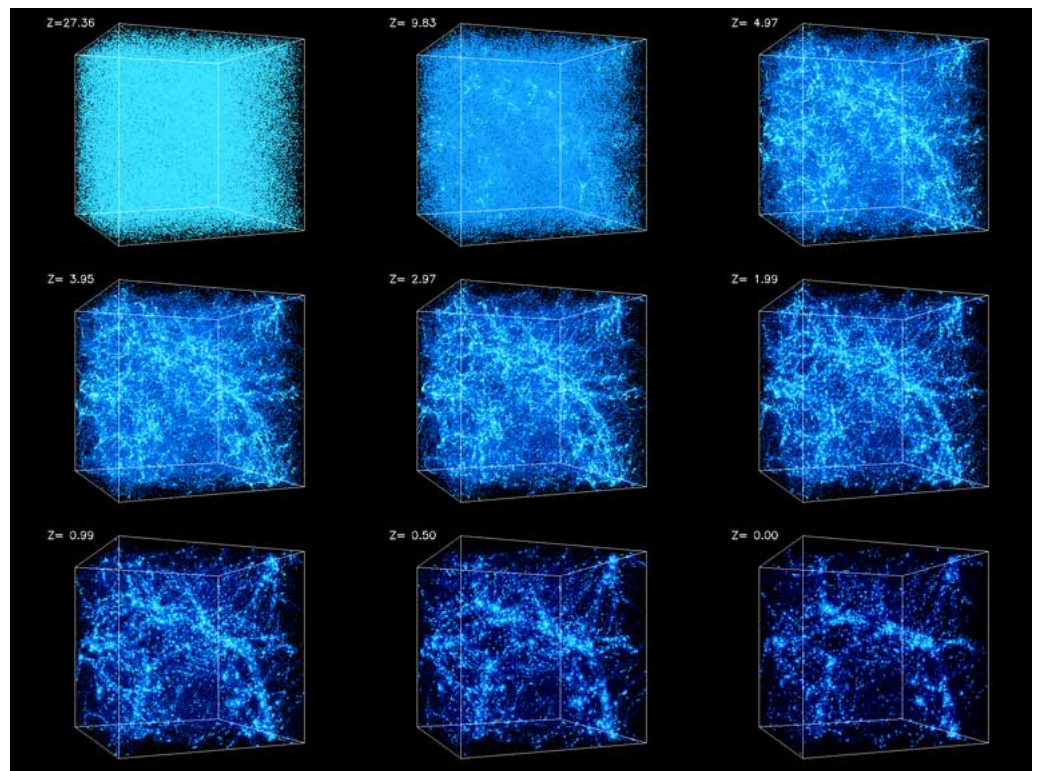


Figure 29.2 Computer simulations of evolution of large-scale structure of universe, beginning with the nearly smooth, early universe at redshift $z = 27.36$ (upper left) to the extensive structure in the current local universe at $z = 0$ (lower right).

Part V

Cosmology

30 Newtonian Dynamical Model of Universe Expansion

30.1 Critical Density

In its observational form, Hubble's law relates the redshift z of galaxies to their distance d ,

$$z = H_o d / c, \quad (30.1)$$

where c is the speed of light, and the Hubble constant H_o has units of inverse time. For nearby galaxies, the Doppler formula implies that the redshift is just linearly proportional to the speed of recession v ,

$$z = \frac{\Delta\lambda}{\lambda} = \frac{v}{c}, \quad (30.2)$$

which when applied to eqn. (30.1) gives the velocity form of Hubble's law,

$$v = H_o d. \quad (30.3)$$

This form has the simple and obvious interpretation that we currently live in an expanding universe. Indeed, if H_o is strictly taken to be constant, then its inverse defines the "Hubble time" (see equation (22.2)),

$$t_H \equiv \frac{1}{H_o} \approx \frac{10 \text{ Gyr}}{h_o} \quad ; \quad h_o \equiv \frac{H_o}{100(\text{km/s})/\text{Mpc}}, \quad (30.4)$$

which effectively marks the time in the past since the expansion began. As such, this Hubble time provides a simple estimate of the age of the universe since the "Big Bang", with the latter equality giving the age in Gyr in terms of the scaled Hubble parameter $h_o \equiv H_o/(100 \text{ (km/s)/Mpc})$.

But more realistically, one would expect the universe expansion to be slowed by the persistent inward pull of gravity from its matter, much the way that an object launched upward from Earth is slowed by its gravity. Indeed, a key question is whether gravity might be strong enough to stop and even reverse the expansion, much as occurs when an object is launched with less than Earth's escape speed.

For two points separated by a distance $d = r$, the relative speed is set by the Hubble law $v = H_o r$. The associated kinetic energy-per-unit-mass associated with the universe's expansion is thus

$$KE = \frac{v^2}{2} = \frac{H_o^2 r^2}{2}. \quad (30.5)$$

For a uniform density ρ , the total mass in a sphere of radius r centered on the other point is just $M(r) = 4\pi r^3 \rho / 3$. The associated gravitational potential energy-per-unit-mass is thus

$$PE = \frac{GM(r)}{r} = \frac{4\pi}{3} G \rho r^2. \quad (30.6)$$

Setting $KE = PE$, we can readily solve for the present-day critical density needed to just barely halt the expansion,

$$\rho_{co} = \frac{3H_o^2}{8\pi G} = 1.87 \times 10^{-29} h_o^2 \approx 9.2 \times 10^{-30} \frac{\text{g}}{\text{cm}^3} ; \quad H_o \approx 70(\text{km/s})/\text{Mpc}, \quad (30.7)$$

The last evaluation applies for the current observationally inferred, best value of the Hubble constant, $H_o \approx 70 (\text{km/s})/\text{Mpc}$, i.e., $h_o = 0.7$. Note that the arbitrary distance r has cancelled out, demonstrating that this critical-density condition (30.7) applies to the expansion as a whole. If the universe has a present-day density $\rho_o > \rho_{co}$, the expansion will be stopped and even reversed, as we will now quantify by solving for the level of this gravitational deceleration.

30.2 Gravitational deceleration of increasing scale factor

Building upon this notion of gravitationally induced slowing of a critically expanding universe, let us now consider the net deceleration for a universe with a *non-critical* density ρ that is still uniform in space, but changes in time due to the expansion. Writing the present-day distance as $d = r(t = 0) \equiv r_o$ and present-day density as $\rho_o \equiv \rho(t = 0)$, then since volume changes with expansion radius as r^3 , we can see from mass conservation that the density at other times must scale as $\rho(t) = \rho_o r_o^3 / r(t)^3$. The self-gravity of this mass density then causes a *deceleration* of the expansion,

$$\ddot{r}(t) = -\frac{GM(r)}{r^2} = -\frac{4\pi}{3} G \rho r = -\frac{4\pi G \rho_o r_o^3}{3r^2}, \quad (30.8)$$

where the dots represent time differentiation.

For convenience, let us next introduce a changing spatial *scale factor* for this universal expansion,

$$R(t) \equiv \frac{r(t)}{r_o}, \quad (30.9)$$

so that, by definition $R_o \equiv R(t = 0) = 1$. Hubble's law then gives for the present-day expansion rate,

$$\dot{R}_o \equiv \dot{R}(t = 0) = \frac{\dot{r}(t = 0)}{r_o} = \frac{v}{d} = H_o. \quad (30.10)$$

The deceleration equation (30.8) can thereby be written in the scaled form,

$$\ddot{R}(t) = -\frac{4\pi G \rho_o}{3R^2} = -\frac{\rho_o}{\rho_{co}} \frac{H_o^2}{2R^2} = -\Omega_m \frac{H_o^2}{2R^2}, \quad (30.11)$$

where the very last equality defines the critical-density mass¹ fraction in the present universe,

$$\Omega_m \equiv \frac{\rho_o}{\rho_{co}}. \quad (30.12)$$

Multiplying both sides by the expansion rate $\dot{R}(t)$, we can obtain a first integral of (30.11),

$$\dot{R}^2 = \frac{\Omega_m H_o^2}{R} - k = \frac{\Omega_m H_o^2}{R} + (1 - \Omega_m) H_o^2, \quad (30.13)$$

where k is an integration constant, evaluated in the latter equality by using equation (30.10). Noting that the Hubble constant here provides the *scale* for the time derivative, we can simplify the notation by **measuring time in units of the Hubble time**, and so making the substitution $t/t_H = H_o t \rightarrow t$. In such “Hubble units”, (30.13) takes the simpler form with the Hubble constant replaced by unity ($H_o \equiv 1$),

$$\dot{R}^2 = \frac{\Omega_m}{R} + (1 - \Omega_m). \quad (30.14)$$

Note then that, in addition to our original definition $R_o \equiv R(t=0) = 1$, we now also have, in these units, $H_o \equiv \dot{R}_o \equiv \dot{R}(t=0) = 1$.

The behavior of the expansion solution $R(t)$ depends on the critical density fraction Ω_m , as delineated in the following subsections, and plotted² in figure 30.1. These solutions $R(t)$ vs. t are computed by *inverting* the integral function for the time,

$$t(R) = \int_1^R \frac{dr}{\sqrt{\Omega_m/r + 1 - \Omega_m}} \quad (30.15)$$

where the lower bound of the integral at $R(t=0) = 1$ was chosen so that this time is measured from the present $t=0$, with any smaller $R < 1$ thus occurring in the *past*, $t < 0$. Eqn. (30.15) can be integrated analytically, but except for some special cases noted below, the full mathematical forms are quite complicated, and thus not obviously very instructive.

Empty Universe, $\Omega_m = 0$

The simplest case is that of an “empty” universe, $\Omega_m=0$, representing the limit in which the mass density is too small to induce much gravitational deceleration. We then find that the expansion rate is constant, with $\dot{R} = 1$, which can be readily integrated, together with the boundary condition $R(t=0) = 1$, to give a uniformly expanding scale factor that just increases linearly with time,

$$R(t) = 1 + t. \quad (30.16)$$

This case is illustrated by the straight black line in figure 30.1.

¹ Note that this includes the total mass contributing to gravitational attraction, including both ordinary, baryonic matter, as well as dark matter.

² The relevant solutions here are those with no “cosmological constant” term, $\Omega_\Lambda = 0$; see § 31 below for the meaning of models with $\Omega_\Lambda n_e 0$.

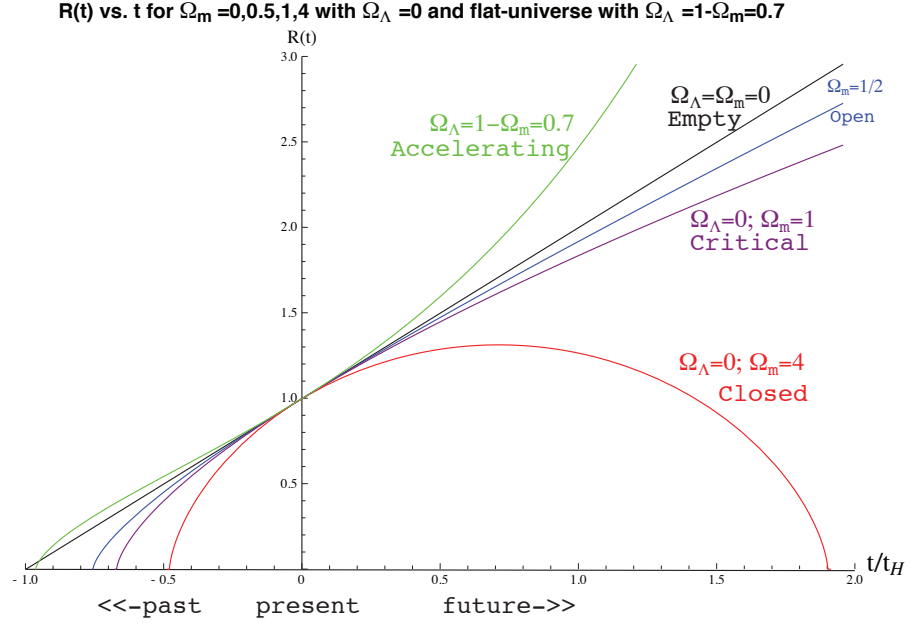


Figure 30.1 Cosmological scale factor R plotted vs. time t in units of the Hubble time $t_H \equiv 1/H_0$, ranging from past ($t < 0$), through present ($t = 0$), to future ($t > 0$), for various combinations for matter critical density fraction Ω_m and cosmological constant energy density fraction Ω_Λ .

30.2.1 Critical Universe, $\Omega_m = 1$

Another case allowing simple integration is that of a critically dense universe, $\Omega_m = 1$, for which (30.14) gives the expansion rate,

$$\dot{R} = R^{-1/2}. \quad (30.17)$$

Upon integration with the boundary condition $R(t = 0) = 1$, this gives the solution

$$R(t) = \left(1 + \frac{3}{2}t\right)^{2/3}. \quad (30.18)$$

This solution thus still expands forever, but approaches a vanishing rate, $\dot{R} \rightarrow 0$ as $t \rightarrow \infty$. It is illustrated by the purple curve in figure 30.1.

30.2.2 Closed Universe, $\Omega_m > 1$

For a still-higher density fraction, $\Omega_m > 1$, the self-gravity can *halt* and *reverse* the expansion. From (30.14) the zero expansion rate $\dot{R} = 0$ occurs at a maximum

scale factor,

$$R_{max} = \frac{\Omega_m}{\Omega_m - 1}. \quad (30.19)$$

As the universe thus eventually closes back on itself, this is known as a “closed” universe. It is illustrated by the red curve in figure 30.1.

30.2.3 Open Universe, $\Omega_m < 1$

Finally, for subcritical density, the expansion again continues forever, but now with a non-zero asymptotic rate, given by taking $R \rightarrow \infty$ in (30.14),

$$\dot{R}_\infty = \sqrt{1 - \Omega_m}, \quad (30.20)$$

which implies that today’s Hubble constant H_o would shrink by this factor $\sqrt{1 - \Omega_m}$ in the distant future. This is known as an “open” universe, illustrated by the blue curve in figure 30.1.

30.3 Redshift vs. distance: Hubble law for various expansion models

Let us next consider how these various theoretical models for the universe connect with the observable redshift that indicates its expansion. Up to now, we’ve considered this redshift to be the result of the Doppler effect associated with distant galaxies receding from us at a speed that increases with distance, giving the speed-distance form of the Hubble law (30.3).

But an alternative, indeed more general and physically more appropriate perspective, is that this redshift is actually just a consequence of the *expansion of space itself!*

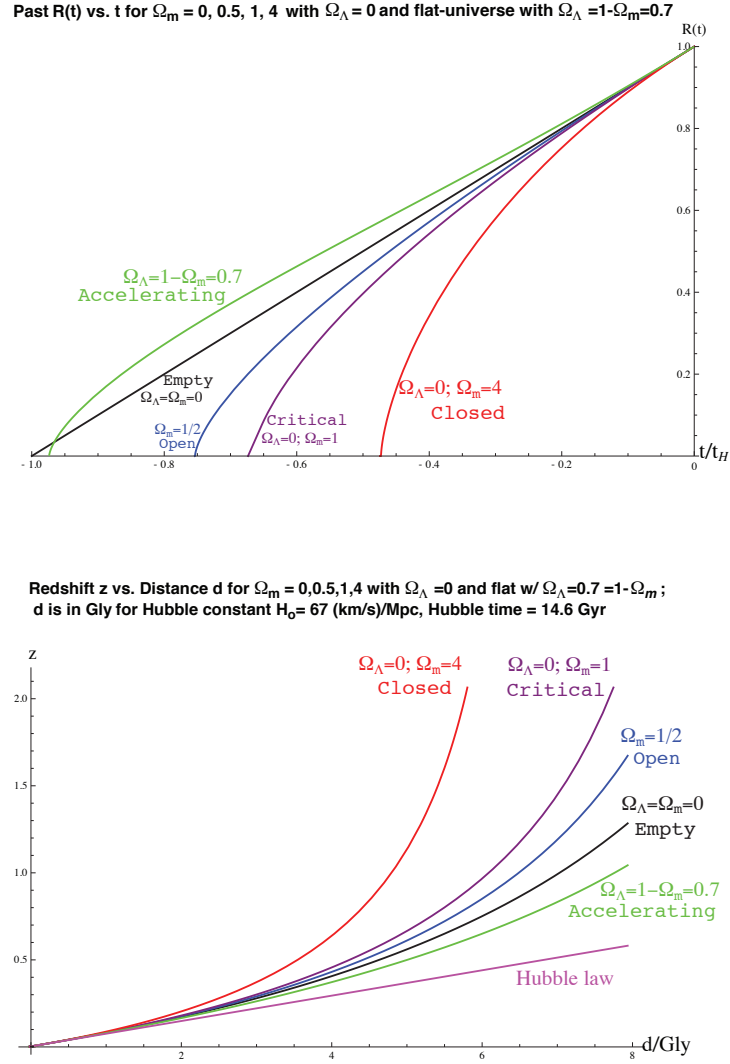
Recall that the basic definition of redshift is given in terms of the difference $\Delta\lambda = \lambda_{obs} - \lambda_{em}$ between the *observed* wavelength λ_{obs} and the originally *emitted* wavelength λ_{em} ,

$$z(d) \equiv \frac{\Delta\lambda}{\lambda_{em}} = \frac{\lambda_{obs}}{\lambda_{em}} - 1 = \frac{1}{R(t = -d/c)} - 1. \quad (30.21)$$

The last equality here follows directly from the definition of the scale factor R as the ratio of a length (here the emitted wavelength) at some remote time (set here by the light travel time $t = -d/c$ to the emitting object at distance d) to that observed at the present time.

For past times that are small compared to the Hubble time, $-t \ll 1/H_o$, Taylor expansion gives $R(t) \approx R(t = 0) + \dot{R}_o t = 1 - H_o d/c$. When applied to (30.21), with further first-order expansion of the inverse binomial, this gives a simple linear Hubble law for distances small compared to Hubble distance $d_H \equiv c/H_o$,

$$z(d) \approx \frac{1}{1 - H_o d/c} - 1 \approx \frac{H_o d}{c} = \frac{d}{d_H} \quad ; \quad d \ll d_H \equiv c/H_o, \quad (30.22)$$



S

Figure 30.2 *Top:* Same as figure 30.1, but focusing only on past times, $t < 0$. *Bottom:* Associated observable redshift z vs. distance d , measured in Giga-light-years (Gly), assuming the current best estimate for Hubble constant $H_0 \approx 70$ (km/s)/Mpc, giving a Hubble time, $t_H = 1/H_0 = 14.6$ Gyr.

thus recovering the standard linear Hubble law (30.1).

But because the redshift depends on the *inverse* of the scale factor, for distances that are not small compared to the Hubble distance, the redshift-vs.-

distance relation becomes distinctly *nonlinear*, even for the linear expansion $R = 1 - d/d_H$ solution that applies for an empty universe with $\Omega_m = 0$. (See black curve in lower panel of figure 30.2.)

For the same selection of expansion models as in figure 30.1, figure 30.2 compares plots of the scale factor R for past times $t < 0$ (top) to the associated variation of redshift z vs. distance d (bottom). Note that, as implied by the expansion (30.22), all the models converge to the simple linear Hubble law (purple line) at modest distances, $d \ll d_H = c/H_o$. But for the inferred Hubble constant $H_o \approx 70$ (km/s)/Mpc, giving a Hubble time $t_H \approx 14.6$ Gyr – which sets the *slope* of that initial line –, we see that at distances beyond 1-2 Gly, these models each start to deviate significantly from this linear Hubble law.

This deviation is greatest for the closed universe case, but because of the inverse relation between redshift z and scale factor R , even the case of an empty universe ($\Omega_m = 0$), with constant rate of expansion ($\dot{R}(t) = H_o$), shows a substantial deviation from the linear Hubble law for distances beyond about 2 Gly.

30.4 Questions and Exercises

Quick Question 1:

- (a) What is the age (in Gyr) of an “empty” universe with constant expansion and Hubble constant $H_o = 67$ (km/s)/Mpc?
- (b) What is the age (in Gyr) of a “critical” universe ($\Omega_m = 1$) and Hubble constant $H_o = 67$ (km/s)/Mpc?

Exercise 29-1: Critical universe redshift.

Consider a critical universe $\Omega_m = 1$ without dark energy ($\Omega_\Lambda = 0$) and a local Hubble constant equal to the currently inferred best value $H_o \approx 70$ (km/s)/Mpc.

- a. Derive a formula for redshift z vs. distance d (in Mpc).
- b. Show that for small distances $d \ll c/H_o$, this recovers the simple linear Hubble law $cz = H_o d$.
- c. Compute the time since the Big Bang, in Gyr.
- d. Compare this time to the age of a Globular cluster with a main-sequence turnoff at luminosity $L_{to} = 0.75 L_\odot$.
- e. What does this say about the viability of this as a model for our universe? What about closed-universe models with $\Omega_m > 1$? (Assume the above Hubble constant measurement is accurate, and that there is no dark energy.)

Exercise 29-2 Empty Universe:

Next consider the case of an effectively “empty” universe with $\Omega_m = \Omega_\Lambda = 0$, that is again expanding with a locally measured Hubble constant $H_o \approx 70$ (km/s)/Mpc.

- a-d. Repeat parts a-d of Exercise 24-1 for this case of an empty universe.
- e. What does the result in part d here say about the formal viability of this as a model for our universe?

Exercise 29-3: Empty vs. Critical Universe:

- a. For the empty universe model of Exercise 24-2, invert the formula for $z(d)$ to derive

an expression for distance as a function of redshift z . For this use the notation $d_0(z)$, where the subscript “0” denotes the null value of Ω_m .

b. If a distance measurement is accurate to 10%, at what minimum redshift z_o can one observationally distinguish the redshift vs. distance of an empty universe from a strictly *linear* Hubble law $d = cz/H_o$.

c. Using the results from Exercise 24-1a, now derive an analogous distance vs. redshift formula $d_1(z)$ for the critical universe with $\Omega_m = 1$ (and $\Omega_\Lambda = 0$).

d. Again if a distance measurement is accurate to 10%, at what minimum redshift z_1 can one observationally distinguish the redshift vs. distance of such a critical universe from a strictly linear Hubble law.

e. Finally, again with a distance measurement accurate to 10%, at what minimum redshift z_{10} can one observationally distinguish the redshift vs. distance of a critical universe from an empty universe?

31 Accelerating Universe with a Cosmological Constant

31.1 White-dwarf supernova as distant standard candles

To test which of these models applies to our universe, one needs to extend redshift measurements to large distances, out to several Gly. As long as an object is bright enough to show detectable spectral lines, measurement of redshift is straightforward, with for example quasars showing redshifts up to $z \approx 6.5$.

But it is much more difficult to get an *independent* measurement of distance for suitably remote objects. The most successful approach has been to use white-dwarf supernovae (WD-SN, a.k.a. type Ia, or SN Ia) as very luminous *standard candles*. Because these supernova all begin with similar initial conditions, triggered when accretion of matter from a companion pushes a white-dwarf star beyond the Chandrasekhar mass limit $M \approx 1.4M_{\odot}$, they tend to have a quite similar peak luminosity, $L \approx 10^{10}L_{\odot}$, corresponding to an absolute magnitude $M \approx -20$. From the observed peak flux F or apparent magnitude m , one can then independently infer the distance $d = \sqrt{L/4\pi F} = 10^{1+(m-M)/5} \text{ pc} = 10^{5+m/5}$. Thus, for example, observational surveys with a limiting magnitude $m \approx +20$ can detect WD-SN out to distance of $d \lesssim 10^9 \text{ pc} = 1 \text{ Gpc}$.

When combined with spectral measurements of the associated redshift z , the data from such white-dwarf supernovae place datapoints in a z -vs.- d diagram like figure 30.2. For modest distances, $d \lesssim 1 - 2 \text{ Gly} < 1 \text{ Gpc}$, the slope of a best-fit line thus provides a direct measurement of the Hubble constant, H_o . But to measure deviations from a linear Hubble law, and so determine which of the above deceleration models best matches the actual universe, there was a concerted effort during the 1990's to discover and observe such supernovae in galaxies at greater and greater distances and redshifts. And as points were added at larger distances, they did indeed show the expected trend above this linear Hubble law, marked by the purple line in figure 30.2.

But in one the greatest surprises of modern astronomy, and indeed of modern science, such data points were found to generally lie *below* the black curve that represents a nearly-empty universe, with a *constant* expansion rate $\dot{R} = H_o$. This immediately *rules out all the decelerating* models that lie above this black curve representing constant-rate expansion.

Instead it implies that the expansion of the universe must be *accelerating*!

Exercise 30-1

- Using the information given in the text, compute the absolute magnitude M at the peak brightness of a type Ia SN.
- Next derive a formula for the associated *apparent* magnitude m as function of distance, measured in Gigaparsec, d_{Gpc} .
- Finally, compute the apparent magnitude of the most remote SN Ia detected so far¹, at $d = 10$ Gly.

31.2 Cosmological Constant and Dark Energy

For the universe's expansion to be accelerating requires that, in opposition to the attractive force of gravity, there must be a positive, repulsive force that pushes galaxies apart. Ironically, in an early (~ 1920) application of his general relativity theory, Einstein had posited just such a universal repulsion term – dubbed the “Cosmological Constant”, and traditionally denoted Λ . This was introduced to balance the attractive force of gravity, and so allow for a static, and thus eternal, model of the universe, which was the preferred paradigm at that time. Then, after Hubble's discovery that the universe is not static but expanding, Einstein completely disavowed this cosmological constant term, famously calling it “his greatest blunder”.

But nowadays, with the modern discovery that this expansion is actually *accelerating*, the notion of something akin to the cosmological constant has been resurrected. The full physical bases and origin are still quite unclear, but the effect is often characterized as a kind pressure or tension of space-time itself, with associated mass-energy density, dubbed “*dark energy*”, parameterized in terms of the fraction Ω_Λ of the critical mass-energy density $\rho_{co}c^2$.

While a rigorous discussion requires a general relativistic treatment beyond the scope of this course, within the above simplified Newtonian model for time evolution of the universe's scale factor R , this dark energy can be heuristically accounted for by adding a *positive* term to the right-side of equation (30.11),

$$\ddot{R}(t) = -\frac{4\pi G\rho_o}{3R^2} + \frac{\Lambda R}{3} = -\Omega_m \frac{H_o^2}{2R^2} + \Omega_\Lambda H_o^2 R. \quad (31.1)$$

Note that now the acceleration transitions from strongly *negative* in the early universe with small scale-factor $R \ll 1$, to strongly *positive* in older universe with large scale-factor $R \gg 1$. The transition, with momentarily zero acceleration ($\ddot{R} = 0$), occurs at a scale factor

$$R_z = \left(\frac{\Omega_m}{2\Omega_\Lambda} \right)^{1/3}. \quad (31.2)$$

Again using the \dot{R} integrating factor and setting $H_o \equiv 1$ to define time in terms

¹ see <http://www.space.com/19198-most-distant-supernova-hubble-discovery-aas221.html>

of the Hubble time, we obtain a generalized first integral solution (cf. equation 30.14),

$$\dot{R}^2 = \frac{\Omega_m}{R} + \Omega_\Lambda R^2 + (1 - \Omega_m - \Omega_\Lambda), \quad (31.3)$$

where we have again evaluated the integration constant by using the boundary conditions $R_o = \dot{R}_o = 1$.

In general relativity gravity is described in terms of the warping, or *curvature*, of space-time². In its application to cosmology, the value of the term in parentheses in (31.3) sets the overall curvature of the whole universe, with positive, negative, and zero values corresponding to curvatures that are similarly positive (like a sphere), negative (like a saddle), and zero (like a flat sheet). Figure 31.1 illustrates these cases.

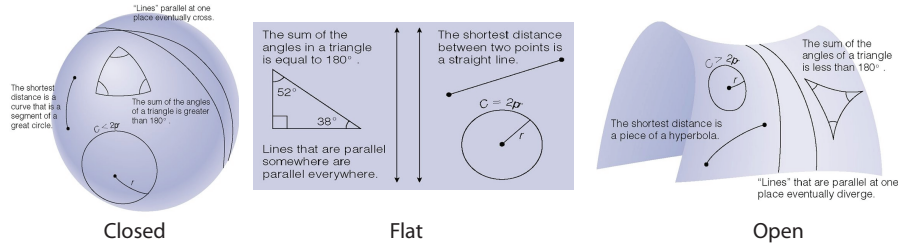


Figure 31.1 Illustration of 3 cases for curvature in ordinary 3D space, ranging from the positive curvature of a closed sphere, to the zero curvature of a flat surface, to the negative curvature with an open saddle. The annotations show how the different geometries lead to different properties for angles and distances.

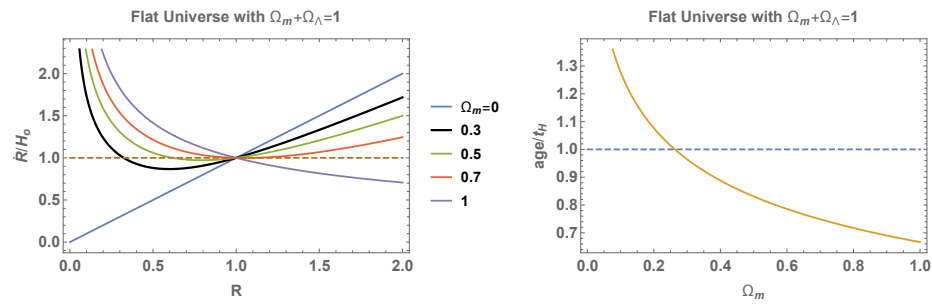


Figure 31.2 *Left:* Expansion rate \dot{R} (in units of H_o) vs. scale factor R for a flat universe with the various labeled values of $\Omega_m = 1 - \Omega_\Lambda$. *Right:* The age of a flat universe (in Hubble time $t_H \equiv 1/H_o$) plotted versus $\Omega_m = 1 - \Omega_\Lambda$.

² In relativity theory, space and time are combined into a coupled *space-time*.

31.3 Flat Universe with Dark Energy

As discussed below, there are strong theoretical arguments (e.g., from the theory of inflation; see §34.2) that the universe must be very nearly *flat*, meaning then that the parentheses term in (31.3) is very nearly zero. This in turn implies that the total energy density is very near the critical value, with $\Omega_m + \Omega_\Lambda = 1$. Using this to eliminate Ω_Λ , we can again cast the range of possible models in terms of the single parameter Ω_m , with (31.3) reducing to

$$\dot{R}^2 = \frac{\Omega_m}{R} + (1 - \Omega_m)R^2. \quad (31.4)$$

For a flat universe with the various labeled values of $\Omega_m = 1 - \Omega_\Lambda$, the left panel of figure 31.2 plots the expansion rate \dot{R} (in units of Hubble constant H_o) vs. scale factor R . Note that for $\Omega_m > 0$, the expansion starts very rapidly, but then declines to a minimum of order H_o , and finally increases again for large R . The bold black curve is for the case $\Omega_m = 0.3$, which as we will discuss below, is roughly the best-fit value for our universe. Note that for much of the evolution of $0.2 \lesssim R < 1$, the model has $\dot{R} = H_o$. The right panel shows that the associated age for such a case with $\Omega_m = 0.3$ is very close to Hubble time $t_H = 1/H_o$ that applies for the simple case of a constantly expanding universe $R(t) = 1 + H_o t$.

31.3.1 Exponential expansion of flat, matter-empty universe

A simple sample for the full solution is again for the case of a matter-empty universe, $\Omega_m = 0$, for which we find $\dot{R} = \pm R$. Choosing the plus root to represent the observed case of expansion, we find

$$R(t) = e^{H_o t} = e^{t/t_H}. \quad (31.5)$$

Thus, in contrast to the previous case of constant expansion for an empty universe with $\Omega_m = \Omega_\Lambda = 0$, for a dark-energy-dominated, flat universe with $\Omega_\Lambda = 1$, the expansion actually *accelerates exponentially*, with an e-fold increase each Hubble time!

31.3.2 General solutions for flat universe with dark energy

In fact, even for the more general case with $0 < \Omega_m < 1$, note that as R increases, the rate again becomes dominated by the second (cosmological acceleration) term in (31.4), implying $\dot{R} \sim +R$ and thus again an exponential expansion at large times. The full solution can again be obtained by inverting the time integral solution, now given by (cf. equation 30.15)

$$t(R) = \int_1^R \frac{dr}{\sqrt{\Omega_m/r + (1 - \Omega_m)r^2}}. \quad (31.6)$$

Again, this integral has analytic solutions, but the forms are complex and so not very insightful to display. But if we extend the upper bound to $R = 0$, we can

obtain a simple analytic form for the universe’s age $t_a \equiv t(R = 0)$ as a function of Ω_m ,

$$\frac{t_a}{t_H} = \frac{\ln \left(\frac{2}{\Omega_m} + \frac{2\sqrt{1-\Omega_m}}{\Omega_m} - 1 \right)}{3\sqrt{1-\Omega_m}}. \quad (31.7)$$

The right panel of figure 31.2 plots t_a/t_H vs. Ω_m . Note that for $\Omega_m = 0.3$, the associated age, $t_a = 0.964t_H$, is very close to Hubble time $t_H = 1/H_o$ that applies for the simple case of a constantly expanding universe $R(t) = 1 + H_o t$.

The green curves in figures 30.1 and 30.2 plot the solution for this $\Omega_m = 0.3$ and thus $\Omega_\Lambda = 0.7$, which turns out to best fit the SN data (as well as other constraints from fluctuations in the Cosmic Microwave Background, CMB). This implies that the combination of ordinary and dark matter makes only about $\sim 30\%$ of the mass-energy density of the universe, with the other $\sim 70\%$ in the form of this mysterious dark energy!

Inspection of figures 30.1 and 30.2 shows that the green curves for this dark-energy model are actually not too different from the basic black curves, which represent the very simple “empty universe” model without *any* kind of matter-energy. In the absence of any forces, this model gives a simple “coasting” solution, with scale factor $R(t) = 1 + H_o t$. Its rough agreement with the dark-energy model means the dark energy can be roughly thought of as providing an outward pressure that approximately cancels the inward attraction from gravity. Since the net force is nearly zero, the dark-energy solution is also nearly coasting, $R(t) \approx 1 + H_o t$, at least for the universe *up to its present age*.

But recall that the first term on the RHS of (31.4), which represent in the inward pull of gravity, declines as $1/R$ as the scale factor R gets large, whereas the second term, representing the cosmological constant, actually *increases quadratically* with increasing R . Thus quite unlike a truly empty, coasting universe, for which the scale factor just *increases linearly* in time, $R(t) = 1 + H_o t$, a universe with a non-zero cosmological constant $\Lambda > 0$ will eventually grow *exponentially*, increasing by an e-fold every Hubble time $t_H = 1/H_o$.

31.4 The “Flatness” problem

One general puzzle for any model of the universe is that having the universe be nearly flat today, with total $\Omega_o = \Omega_m + \Omega_\Lambda \approx 1$, requires that it must have been even *much* flatter, with $\Omega(t)$ much closer to unity, in the past ($t < 0$).

To see this, let us write the constant total energy-per-unit-mass E_{tot} of the expanding universe in terms of the sum of its associated kinetic and potential energy components,

$$v^2 - \frac{2GM(r)}{r} = H^2 r^2 - \frac{8\pi G \rho r^2}{3} = 2E_{tot}. \quad (31.8)$$

By dividing by the second term in the middle expression, this can be recast into

the form

$$\frac{1 - \Omega(t)}{\Omega(t)} = \frac{\rho_{co}}{R(t)^2 \rho(t)} \frac{1 - \Omega_o}{\Omega_o} \quad (31.9)$$

where $\Omega(t) \equiv \rho(t)/\rho_c(t)$ is the critical density fraction at some earlier time t , with $\rho_c(t) \equiv 3H(t)^2/8\pi G$ a generalization of equation (30.7) to define the critical density at this time when the Hubble constant is $H(t)$. On the right-hand-side, the total energy and other constants have thus been cast in terms the critical density Ω_o in the current-day universe. If this Ω_o differs from unity by some small fraction, say $|1 - \Omega_o| \approx 0.01$ (i.e. 1%) in the current-day universe, then in the earlier universe, the difference is smaller by a factor

$$|1 - \Omega(t)| \approx \frac{0.01}{R(t)^2 \rho(t)/\rho_{co}} \approx 0.01 R(t) \quad ; \quad 10^{-4} < R < 1 \quad (31.10)$$

$$\approx 100 R^2 \quad ; \quad R < 10^{-4}. \quad (31.11)$$

The upper equality assumes a matter-dominated universe with density $\rho \sim 1/R^3$. But as, discussed below (see §§32.1 and 33.1), the *temperature* of the universe scales as $T \sim 1/R$; thus, in the early universe with $R < 10^{-4}$, the energy density was dominated by *radiation*, since radiation's energy density scales as $U_{rad} \sim T^4 \sim 1/R^4$, i.e. one higher factor of $1/R$ than the $\rho \sim 1/R^3$ scaling of matter. Extending back to very early times, we thus require $|1 - \Omega| \sim R^2 \rightarrow 0$, meaning then that any “initial” deviations from flatness had to have been *extremely tiny*.

If instead, the initial Ω had been even slightly above unity, the fledgling universe would have recollapsed as a tiny, closed universe. Alternatively, if Ω had been even slightly below unity, the universe would have expanded at such a high rate that galaxies would not have had time to form. Overall, this required fine-tuning to make $|1 - \Omega|$ initially very small is known as the “flatness” problem for reaching the kind of moderately expanding, mature universe we live in today.

Let us next consider further the temperature history and associated properties of the universe extending back to such early times of a “Hot Big Bang”.

32 The Hot Big Bang

32.1 The temperature history of the universe

The smaller scale factor of the past universe clearly means its overall averaged density was higher than it is today. But what might we conclude about the overall *temperature* history of the universe? In the present-day universe the temperature of individual structures varies widely, e.g. from millions of Kelvin in the interiors of stars, to just a few degrees above absolute zero in cold giant molecular clouds, and so it might seem absurd to even speak of a single temperature for the *whole* universe.

But if we go back in time before all this structure, when the density of the universe was much higher and much smoother, there was a kind of *thermal equilibrium* that led to a quite well-defined characteristic temperature. Intuitively we can expect that in the smaller, more compressed, and thus much denser early universe, the temperature should also be correspondingly much higher.

And indeed, as discussed below (see equation 33.1), it turns out that the temperature of the early universe scaled *inversely* with the scale factor, $T(t) \sim 1/R(t)$, which also means that it increases linearly with the associated redshift, $T(z) \sim 1 + z$. Figure 32.1 illustrates the overall temperature history of the universe extending to very early times, with very high redshifts and very high temperatures.

For example, at a redshift of $z \approx 1000$, corresponding to a scale factor $R \approx 10^{-3}$, it turns out the temperature of the universe was about as hot as the surface of a relatively cool star, $T \approx 3000 \text{ K} \approx T_{\odot}/2$. And much as a star, this hotter early universe emitted radiation according to the Black-Body function $B_{\lambda}(T)$ for that temperature, with an original emitted spectrum that had its peak at a wavelength $\lambda_{max} = 500 \text{ nm } T_{\odot}/T \approx 1 \mu\text{m}$.

But in the present-day universe this radiation should be *redshifted* by a factor $z = 1/R - 1 \approx 10^3$, with a corresponding peak wavelength in the *microwave* region (like in your microwave oven), $\lambda_{max} \approx 10^3 \mu\text{m} \approx 1 \text{ mm}$. Moreover, in contrast to the directed “outward” emission from a star, this cosmic radiation was emitted *isotropically* (equal in all directions), and so would be observed today from all directions in the sky, as what is known as the *Cosmic Microwave Background* (CMB).

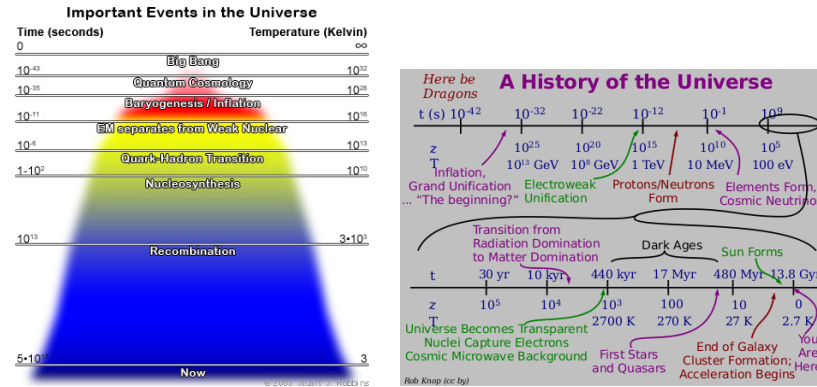


Figure 32.1 Two renditions of key events and eras of the Hot Big Bang, extending back to very early times with very high redshifts and very high temperatures.

32.2 Discovery of the Cosmic Microwave Background (CMB)

Early proponents of this “Hot-Big-Bang” model – most notably Robert Dicke of Princeton – actually predicted such a CMB before it was detected, rather serendipitously, in 1965 by two engineers named Penzias and Wilson from Bell Labs. They were actually just trying to reduce the persistent noise that was inherent in the radio receivers they were developing for communications, in some ways the predecessors of microwave antennae used for cell phones today. After working hard to reduce electronic and other¹ possible sources of static, they eventually concluded the noise was actually coming from the sky. Noting moreover that it was constant over both night and day, with a uniform brightness over the whole sky (and not, for example, concentrated along the equator, ecliptic or the plane of the Milky Way), they, with some help from reading an unpublished preprint by Dicke and his colleagues, identified it as the predicted CMB. This momentous discovery, which provided striking confirmation of the Hot Big Bang model, eventually earned them (but not Dicke) a share of the 1978 Nobel Prize in Physics.

Subsequent observations have shown that the CMB is indeed isotropic to a very high precision ($< 10^{-4}$). Moreover, as illustrated in figure 32.2, it also follows both the form and absolute surface brightness² of the Planck Black-Body function to a similarly high precision, with an inferred temperature $T_{cmb} = 2.726 \pm 0.001$ K. This can be considered as the present-day “temperature of our universe”.

¹ including, they reported, from pesky avian deposits of “dielectric material” on the antennae

² Recall that, in contrast to the flux from a localized source, surface brightness of an angularly resolved source does not decline with distance. Thus, once the redshift expansion of the universe is accounted in reduction of the CMB temperature, the surface brightness of the CMB is the same today as what was emitted at the end of the recombinations era!

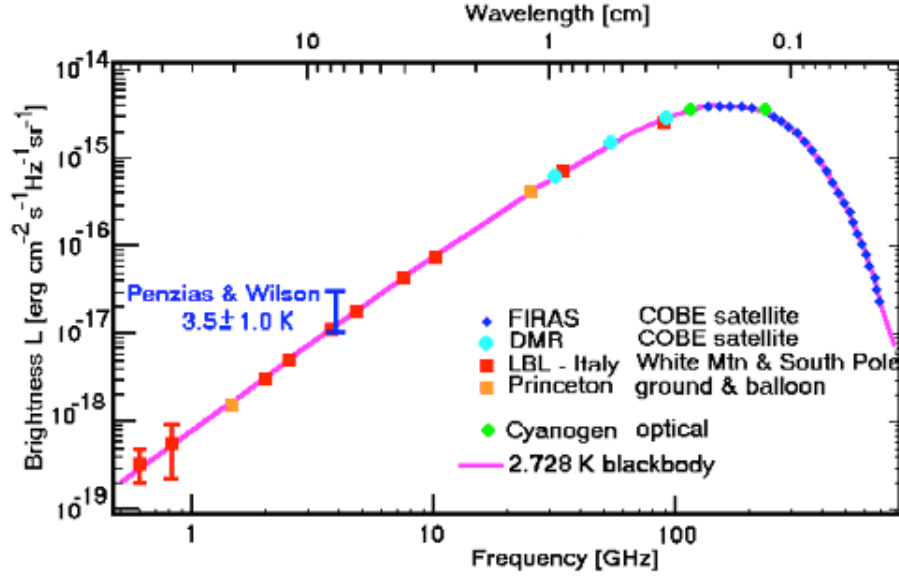


Figure 32.2 Sky brightness of CMB vs. frequency (bottom axis) or wavelength (top axis) on a log-log scale, showing the nearly perfect fit of data from COBE and other measurements to a Planck Black-Body function of temperature $T_{cmb} = 2.728$ K (purple curve).

32.3 Fluctuation Maps from COBE, WMAP, Planck

Although the CMB appears isotropic and uniform down to levels $< 10^{-4}$, the universe we live in today is very non-uniform, with large-scale structure, superclusters, galaxies, stars, and planets. Even with the extra mass from dark matter to enhance the mutual gravitational attraction, any contraction to form this extensive structure still requires initial “seeds” in the form of small-amplitude fluctuations in local density. From simulation models for the formation of large-scale structure, it was predicted during the 1980’s that the level of fluctuations needed would impart small fluctuations in the CMB at the level of a few part per one-hundred thousand, i.e. a few times 10^{-5} , implying temperature fluctuations up to $\Delta T \lesssim 10^{-4} T_{cmb} \approx 300 \mu\text{K}$.

Detecting such fluctuations thus became a major goal for observation and experiment. For ground-based observations it is very difficult to remove the effects of Earth’s atmosphere to a level that doesn’t mask the predicted fluctuations, though there was some success, for example from a balloon-born experiment called *Boomerang* that circled the south pole. But the clearest results came from a series of orbiting satellites named COBE (COsmic Background Explorer, launched in 1989), WMAP (Wilkinson Microwave Anisotropy Probe, launched in 2001), and Planck (launched in 2009). COBE succeeded in measuring fluctuations at a level of about $200 \mu\text{K}$, or $\lesssim 10^{-4}$, but its resolution was limited to large

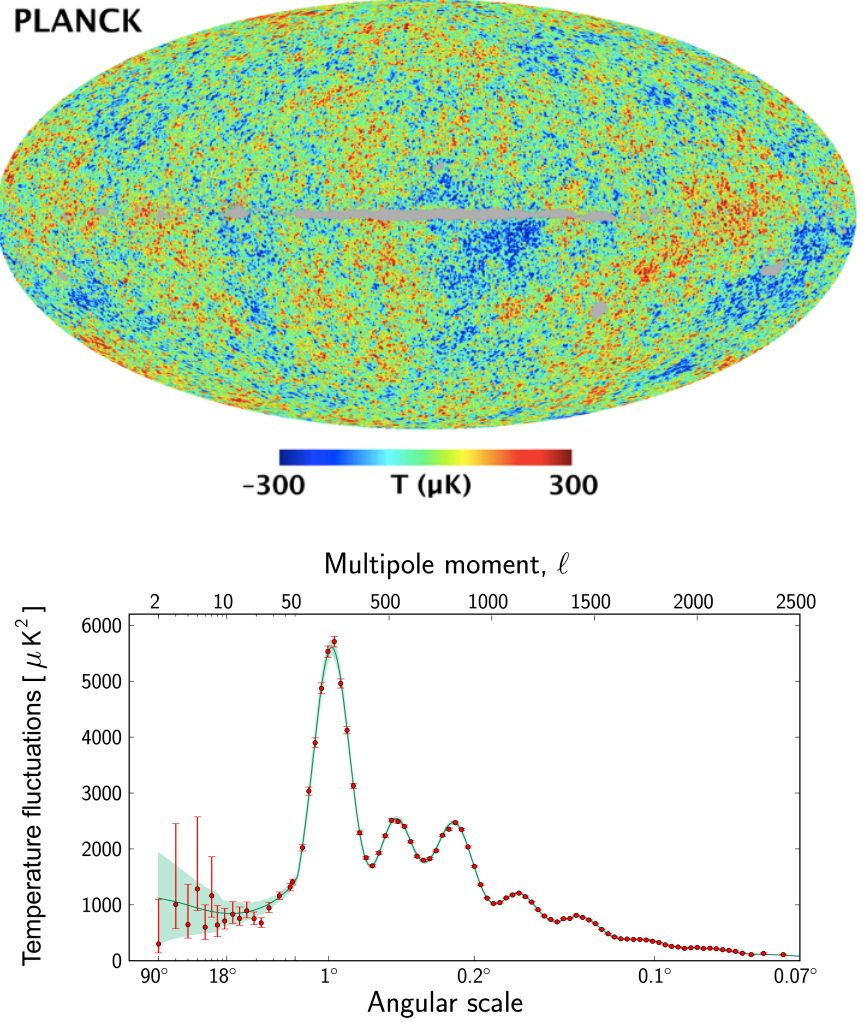


Figure 32.3 *Top:* Sky map of temperature fluctuations in the Cosmic Microwave Background (CMB), as measured by the Planck satellite. *Bottom:* Power spectrum of temperature fluctuations plotted vs. angular scale (lower axis) or spherical harmonic multipole moment ℓ (upper axis).

angular scales, $> 7^\circ$. WMAP, and now Planck, have greatly improved both the precision and the angular resolution, with Planck measuring fluctuations down to a precision of a few μK (i.e., $\Delta T/T \sim 10^{-6}$), at angular scales $< 0.1^\circ$.

The top panel of figure 32.3 shows a full-sky map (in galactic coordinates, with galactic plane extending horizontally from the galactic center) of the CMB temperature fluctuations (in μK), as measured by the Planck satellite. The

color range over $\pm 300 \mu\text{K}$ represents relative fluctuations up to $\pm 300 \mu\text{K} \sim \pm 10^{-4} T_{cmb}$, with red hotter and blue cooler.

The spatial power spectrum in the lower panel shows that these fluctuations occur over a range of angular scales, with main peak at about 1° . While the observed CMB comes from the last scattering during the recombination era, the fluctuations originate from processes before this era. Much as measurement of seismological waves generated in an earthquake provide information on the interior structure of the Earth, these measures of CMB fluctuation power peaks provide information on the pre-recombination evolution of the universe, and place strong constraints for basic cosmological parameters.

Specifically the Planck analysis quotes values $\Omega_b = 0.049$ for the fraction of ordinary (Baryonic) matter, $\Omega_{dm} = 0.268$ for the fraction of dark matter, $\Omega_\Lambda = 0.682$ for the fraction of dark energy, $H_o = 68.15 \text{ km/s/Mpc}$ for the Hubble constant, and 13.82 Gyr for the age of the universe.

Exercise 26-1:

- a. For a Planck function $B_\nu(T)$ at frequency ν for temperature T , show that the fractional distribution of energy in a given frequency interval ν and $\nu + d\nu$ – given by $B_\nu(T)d\nu/B(T)$ (where $B(T) = \sigma_{sb}T^4/\pi$ is the frequency-integrated emission given in equation, 5.1) – depends only on the dimensionless ratio, $h\nu/kT$
- b. Similarly for the wavelength form of the Planck function $B_\lambda(T)$, show that the fractional distribution of energy in wavelength λ depends only on the dimensionless ratio, $hc/\lambda kT$

33 Eras in the Evolution of the Universe

33.1 Matter-dominated vs. Radiation-dominated eras

A key property of the Planck function is that the overall form of fractional energy distribution over wavelength depends only on the *product* λT . Thus the redshift of an observed vs. emitted wavelength – by a factor $\lambda_{obs}/\lambda_{em} = 1 + z$ – can just be accounted for by reducing the observed vs. emitted temperature – by a factor $T_{obs}/T_{em} = 1/(1 + z)$.

But since $1 + z = 1/R$, this then implies that this radiation temperature of the universe just increases with the inverse of the scale factor,

$$T(t) = \frac{T_{cmb}}{R(t)} = T_{cmb} (1 + z). \quad (33.1)$$

Since the energy density of radiation scales as $U(T) = a_{rad}T^4$ (where $a_{rad} \equiv 4\sigma_{sb}/c$), we conclude that the radiative energy density has a scaling $U \sim T^4 \sim 1/R^4$ that is *steeper* (by one factor of $1/R$) than the density scaling, $\rho \sim 1/R^3$, of ordinary matter. For the present-day matter density $\rho_o = \Omega_m \rho_{co}$, the ratio of matter to radiation energy density is

$$\frac{\rho_o c^2}{U(T_{cmb})} = \frac{\Omega_m c^2 (3H_o^2/8\pi G)}{a_{rad} T_{cmb}^4} \approx 4.2 \times 10^4 h^2 \Omega_m \approx 6000, \quad (33.2)$$

where $h \equiv H_o/(100 \text{ km/s/Mpc})$, and the last equality comes from applying the standard values $h \approx 0.7$ and $\Omega_m \approx 0.3$. Thus in our present-day universe *matter dominates over radiation* in terms of the associated mass-energy density.

However, since this ratio declines in direct proportion to the decreasing scale factor R , we find that at a time with $R \approx 1/6000 \approx 10^{-4}$, when the redshift was approximately $z = 1/R - 1 \approx 10^4$, there is a transition to a higher density in radiation than matter, with earlier times with $R < 10^{-4}$ (and so $z > 10^4$) thus representing a *radiation-dominated* era.

Moreover, even though the mass-energy of the present-day universe is dominated by matter over radiation, it turns out that the *number* of CMB photons $n_\gamma(T_{cmb})$ actually greatly exceeds the number n_H of Hydrogen atoms or protons. Since Hydrogen is a mass fraction $X_H \approx 0.73$ of the ordinary matter that only amounts to about 5% of the critical density ρ_{co} , the present-day Hydrogen

number density is about,

$$n_{Ho} \approx \frac{0.73 \times 0.05 \rho_{co}}{m_p} = 1.8 \times 10^{-7} \text{ cm}^{-3}. \quad (33.3)$$

The number of CMB photons can be estimated by dividing the energy density by an average photon energy, which for the CMB temperature is $\langle E \rangle \approx 3kT_{cmb} \approx 7 \times 10^{-4} \text{ eV}$. This gives

$$n_{\gamma o}(T_{cmb}) \approx \frac{U(T_{cmb})}{\langle E \rangle} \approx \frac{a_{rad} T_{cmb}^4}{3kT_{cmb}} = 360 \text{ cm}^{-3} \approx 2 \times 10^9 n_{Ho}, \quad (33.4)$$

which shows that the photon number is more than a billion times the proton density.

Moreover, since $n_{\gamma} \sim U(T)/T \sim T^3 \sim 1/R^3$, this photon density has the same $\rho \sim 1/R^3$ dependence on scale factor as the matter density. As such, the ratio $n_{\gamma}/n_p \sim 10^9$ thus remains roughly *constant* (!) at this high value all the way back to the formation of the CMB, and indeed well into the radiation-dominated era.

As discussed in §33.3, this ratio plays an important role in the relative abundances of He and other light elements that form in the “era of nucleosynthesis”, when $T \approx 10^9 \text{ K}$.

But first, let us next consider more carefully the formation of the CMB, at the time when the temperature was $T \gtrsim 3000 \text{ K}$, cool enough for electrons to recombine with protons, known thus as the “recombination era”.

33.2 The recombination era

In the early epochs of the Hot Big Bang, the temperature was so high that all the Hydrogen was fully ionized, with proton number density $n_p = n_H$. Moreover, if for simplicity we neglect the contributions from Helium, then overall charge neutrality requires an equal number of electrons, and so $n_e = n_p = n_H$. Because electrons can so readily scatter radiation, the photons of this era were efficiently trapped, much as they are in the interior of a star. But as the universe cooled, the protons and electrons recombined to make neutral Hydrogen, which is much less effective in absorbing or scattering radiation. The photons from this *recombination era* thus were suddenly free to propagate through the universe, becoming redshifted by its expansion to form the CMB we observe today.

We can model this CMB formation much as we model the emitted radiation from a star like our Sun. Recall from the Eddington-Barbier relation of §C.2 (see equation C3) that the surface brightness at the center of the solar disk is set by the Planck function at about unit optical depth along that radial (i.e. $\mu = 1$) line of sight, $I_{obs} \approx B(\tau = 1)$. Analogously, the CMB surface brightness emitted at the recombination era can be derived from the electron-scattering optical-depth. Integrating over the path of the photons, traveling at the speed of light c , from

some past time ($t_p < 0$) to the present day ($t = 0$), this optical depth is given by

$$\tau_e(t) = \int_{t_p}^0 \sigma_{Th} n_e(t) c dt = \sigma_{Th} c \int_{t_p}^0 X_e(t) n_H(t) dt = n_{Ho} \sigma_{Th} c \int_{R(t_p)}^0 \frac{X_e(R)}{R^3 dR/dt} dR, \quad (33.5)$$

where σ_{Th} is the Thompson cross-section for electron scattering (see §D.1 and equation D1), $X_e \equiv n_e/n_H$ is the electron fraction, and the last equality converts this to an integral over scale factor R . For the simple linear expansion (empty) universe that roughly fits observations, we have $dR/dt = H_o$, with H_o the present-day Hubble constant. Using $R = 1/(1+z)$, we can then convert this to an integral over redshift z ,

$$\tau_e(z) = \tau_o \int_0^z X_e(z')(1+z') dz', \quad (33.6)$$

where $\tau_o \equiv n_{Ho} \sigma_{Th} c / H_o \approx 0.0017$ sets the overall scale of the optical depth¹, evaluated here for $H_o \approx 70$ km/s/Mpc.

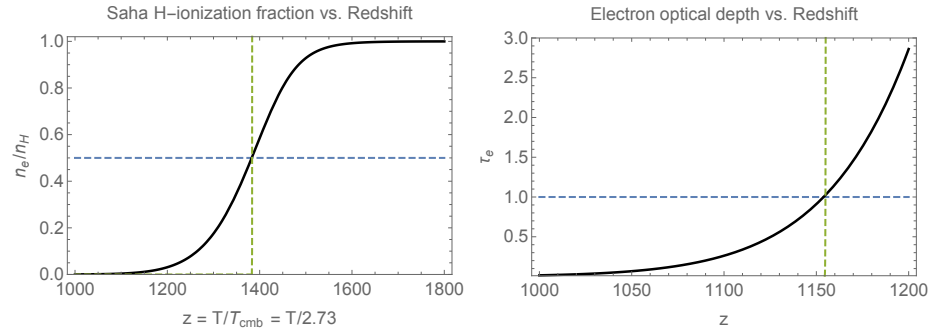


Figure 33.1 *Left:* Electron-to-Hydrogen ratio vs. redshift z , computed from solutions of the Saha-Boltzmann ionization equilibrium equation (33.7) for Hydrogen. *Right:* Electron optical depth τ_e for CMB photons vs. redshift z , computed from equation (33.6).

To proceed, we need to determine the electron fraction X_e . This can be computed from solution of the Saha-Boltzmann ionization equilibrium discussed in §B.2. Applying equation (B4) to the case of pure Hydrogen using $g_1/g_0 \approx 1/2$ with $n_e = n_p$, we can write

$$\frac{X_e^2}{1 - X_e} = \frac{1}{n_H(z)} \left(\frac{2\pi m_e kT(z)}{h^2} \right)^{3/2} e^{-\Delta E_H / kT(z)}, \quad (33.7)$$

where $\Delta E_H = 13.6$ eV is the Hydrogen ionization energy, with $T(z) = T_{cmb}(1+z)$ and $n_H(z) = n_{Ho}(1+z)^3$. Using standard numerical root finding, equation (33.7) can be readily solved to obtain $X_e(z)$, as plotted in the left panel of figure 33.1 for our standard cosmological parameters. The dashed lines show that 50%

¹ If the universe were fully ionized today, τ_o would be the optical depth of the Hubble distance c/H_o .

ionization ($X_e = 0.5$) occurs at a redshift $z_{1/2} \approx 1380$, corresponding to a temperature $T_{1/2} \approx 3700$ K.

The right panel plots the associated redshift variation of the electron optical depth, as computed from equation (33.6). The dashed lines now indicate the level for unit optical depth, $\tau_e(z_{rec}) = 1$, the solution of which gives a recombination era redshift $z_{rec} \approx 1150$, corresponding now to a recombination temperature $T_{rec} \approx 3100$ K. The associated electron fraction $X_e = 0.012$, reflecting the fact that for the much higher density of the recombination era, even a $\sim 1\%$ ionization fraction gives enough free electrons to make the radiation transport marginally optically thick.

These derived values for the redshift and temperature of the recombination agree well with the rough values assumed in the above introduction to the CMB. But they also agree remarkably well with values derived from more complete CMB models.

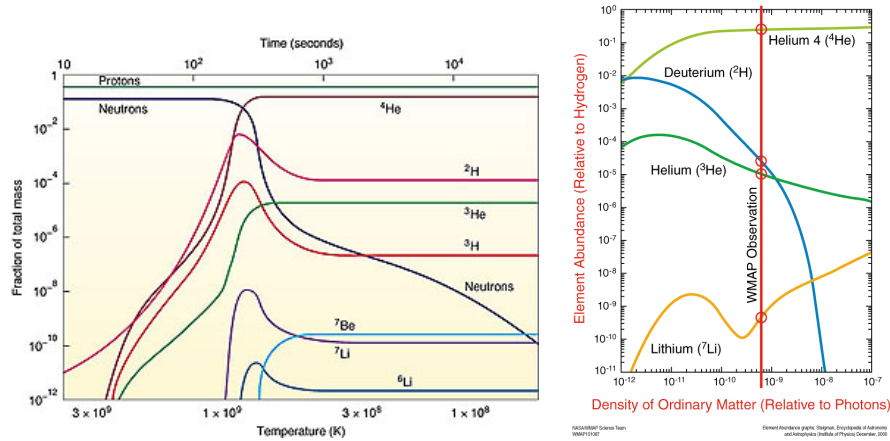


Figure 33.2 *Left:* Relative abundance of various light elements as function of time since the Big Bang (upper axis) or temperature (lower axis). *Right:* The final relative abundances as function of the number density ratio of ordinary (Baryonic) matter relative to photons.

33.3 Era of nucleosynthesis

Another important constraint on the conditions in the early universe, extending to even well before the last scattering surface that formed the CMB, comes from fitting the present-day abundance of Helium and other light elements. While Helium is synthesized in stars, it turns out that most of the Helium in the universe today was actually formed in the first few minutes or so after the Big Bang, when the temperature was several billion degrees (10^9 K). This is called the “era of nucleosynthesis”.

The left panel of figure 33.2 plots the relative abundance of various light elements as function of time or temperature of the Hot Big Bang. Neutrons created at earlier, hotter times had a number ratio of about 1/7 to protons, and at a temperature of about 10^9K most all these were converted into very stable He nuclei, representing then the $\sim 25\%$ mass fraction of primordial He we see in the universe today. Because neutrons and protons can combine without having to overcome electrical repulsion, this He production occurs quite quickly², over just a few minutes!

However, it does proceed through a chain of reactions that first form the rare isotopes ^2H (deuterium) and ^3He , and so at any given time, there are some small fractions of these. As illustrated in the right panel of figure 33.2, the net final relative abundances of these rare isotopes depends sensitively on the number density ratio of ordinary (baryonic) matter relative to photons.

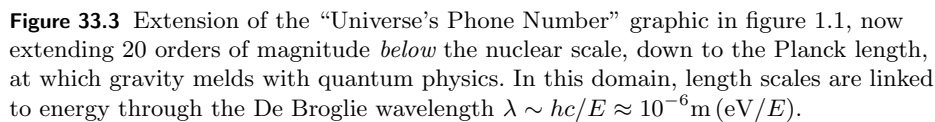
This means we can use present-day measurements of the relative abundances of He and these rarer light elements to put a strong constraint on the matter/photon number ratio, which, as noted in §33.1, stays constant in time. Given the very precisely measured CMB temperature today, we can readily obtain the number density of CMB photons, thus allowing us to convert the inferred matter/photon ratio into a present-day matter density.

The upshot then is that these observed relative abundances of light elements place a strong constraint on the density of baryonic matter, and its associated closure fraction Ω_b , in the present-day universe. In particular, as illustrated by the vertical line in the right panel of figure 33.2, these measurements provide an independent check on the matter density inferred from fluctuations in the CMB measured by WMAP and Planck.

33.4 The particle era

At even earlier epochs, with temperature $T > 10^{10}\text{K}$, it is better to measure temperature in energy units, eV, instead of Kelvin. Recalling that $1\text{ eV} \approx 10^4\text{ K}$, we see that $10^{10}\text{ K} \approx 1\text{ MeV}$, which is about twice the rest mass energy of an electron. At these temperatures, the photons have sufficient energy to *create* pairs of electrons and its antimatter counterpart, the anti-electron, or positron. Reaction of the large number of electrons with protons then make neutrons. As the temperature cools, and the electrons and positron annihilate, the neutron

² Free neutrons are unstable, with a half-life about 15 min, and so are rare in the universe today. As such, present-day production of He in stellar cores requires overcoming the electrical repulsion between two protons, with relatively cool temperature $\sim 10^7\text{ K}$ that only bring the protons within a De Broglie wavelength of each other to allow quantum tunneling. But this proceeds only slowly, requiring a main-sequence lifetime of millions or even billions of years to convert the core H into He. Cores of stars are thus relatively low-temperature “slow cookers” of He compared to the rapid nucleosynthesis in the first few minutes of the Hot Big Bang.



At even higher energies, $T > 10^{13}$ K ≈ 1 GeV, collisions are now above the rest-mass-energy of protons, ~ 1 GeV, and so now create lots of protons + anti-protons. In this particle era, the universe was thus very nearly *symmetric* between matter and anti-matter. But due to quantum fluctuations, for every billion anti-protons, there were about a billion + one protons, from “spontaneous symmetry breaking”. As the temperature cooled, each anti-proton was annihilated with a pairing proton, producing the photons we see in the CMB today, with just the extra one in a billion proton left behind. The upshot is that, because of this spontaneous symmetry breaking of quantum physics, we find ourselves today in a matter universe, instead of an anti-matter universe, with about a billion photons for every proton, a ratio that, as discussed in §33.1, remains to this day.

At even higher temperatures, protons and anti-protons are broken in to sea of “quarks”. These higher temperatures are also associated with a “merging”

of fundamental forces: At $T \sim 250$ GeV, electricity and magnetism merge with the weak nuclear force, giving what is called the *electro-weak* force. Beyond this, it takes a *much* higher energy, $T \sim 10^{16}$ GeV (10^{25} eV $\sim 10^{29}$ K!), to merge the strong force with the electro-weak force. Our best “standard model” for this is called Grand Unified Theory, or GUT, and so this merger is said to occur at the “GUT scale”. By comparison, the most powerful particle accelerator we have on Earth, the Large Hadron Collider (LHC), only reaches $\sim 10^{12}$ eV (maybe extended to 10^{13} eV in the future). This means such particle colliders can’t directly test (or constrain parameters of) the GUT standard model.

The unification of gravity with the GUT force occurs at an even earlier, hotter epoch, known as the Planck scale, with $T \sim 10^{19}$ GeV. There are several competing approaches – e.g. string theory, supergravity – for this unification. But because they operate at such extreme energies that are far beyond what can be tested by collider experiments, they have been developed on purely theoretical and mathematical grounds.

Figure 33.3 extends the cosmic scale range shown in figure 1.1, now adding 20 orders of magnitude below the nuclear scale, down to the Planck length. In this domain, length scales are linked to energy through the De Broglie wavelength, $\lambda \sim hc/E \approx 10^{-6} \text{m} (\text{eV}/E)$. The GUT scale, wherein all the forces except gravity unify, occurs at energy $E_{GUT} \sim 10^{25}$ eV, corresponding to a length scale $\lambda \sim 10^{-31}$, some 16 orders of magnitude below the nuclear scale, but still 4 orders of magnitude above the Planck scale. By comparison, the Large Hadron Collider has just been able to detect the Higgs Boson. With rest-mass energy of 125 GeV, this corresponds to a scale $\sim 10^{-19}$ m, just 4 orders down from the nucleus, and so still 16 orders from the Planck scale. Some wonder if the domain between might not have much of interest, a kind of size scale desert. Building colliders to find even higher-energy particles than the Higgs will be difficult and expensive, but extending our cosmological studies to earlier epochs with higher temperatures could provide key constraints on physics at these tiny scales.

The upshot is that the Planck era is at the very frontier, where of our current physical understanding is untested and breaks down into a “quantum foam”.

33.5 Questions and Exercises

Exercise 1:

According to Planck, a quantum of energy E has a De Broglie wavelength $\lambda = hc/E$. According to Einstein, such a quantum of energy has an associated mass $m = E/c^2$. Finally, according to Schwarzschild, a mass m would become a black hole if confined within a radius $r = 2Gm/c^2$. Setting $r = \lambda$, combine these relations to solve for the associated “Planck length” in terms of h , G , and c . This represents the length scale at which gravity and quantum physics meld together.

34 Cosmic inflation

34.1 Problems for standard Hot Big Bang model

Despite its successes in explaining the CMB and the relative cosmic abundances of Helium and other light elements, it became clear that this standard Hot Big Bang model could not readily explain certain quite general properties of the observed universe. These can be broken down into 3 fundamental problems:

1. *Flatness problem* – Why is/was the total Ω so close to one, implying the universe very “flat”? In other words, why is the total energy of the universe nearly zero (kinetic + vacuum = – gravitation)?
2. *Horizon problem* – Why is the universe so isotropic, given that opposite sides of the sky should have been outside each other’s “light-travel horizon” in the early universe, and thus unable to communicate to establish a common temperature?
3. *Structure problem* – What is the origin of all the structure in the present-day universe, given that it started out so homogeneously? What caused the small fluctuations we’ve now detected in the CMB?

34.2 The era of cosmic inflation

To answer these questions, in 1980 a young MIT physicist named Alan Guth proposed that the very early universe, at a time $< 10^{-32}$ s, experienced a period of extreme, exponential *Inflation*, expanding by a factor 10^{30} (!) over that tiny time-scale!

He speculated that this may have been powered by the energy generated in the freezing out of the GUT force from the electro-weak and strong force, i.e. that it occurred toward the end of the GUT era mentioned above.

This notion of “*Cosmic Inflation*” provides potential answers to all 3 problems:

1. *Flatness problem* – The inflation of the universe’s size by a factor $\sim 10^{30}$ means that any curvature in the pre-inflated universe is greatly reduced, much as the curvature of a sphere is reduced by increasing its radius.
2. *Horizon problem* – Since the pre-inflated universe was so small, sections that would end up at opposite sides of our present-day sky were initially very

close together, within each other's light horizon, and thus could be causally homogenized to the nearly same properties.

3. *Structure problem* – This smallness of the pre-inflated universe also means that, like atoms, nuclei, and elementary particles today, it was subject to “quantum fluctuations” associated with the uncertainty principal. The initially tiny physical scale of these fluctuations was amplified by inflation to much larger structures that we see today in the angular spectrum of fluctuations of the CMB. Those in turn were the seeds that, with the help from the extra gravitational attraction of cold dark matter (CDM), form the large-scale structure of the universe we see today.

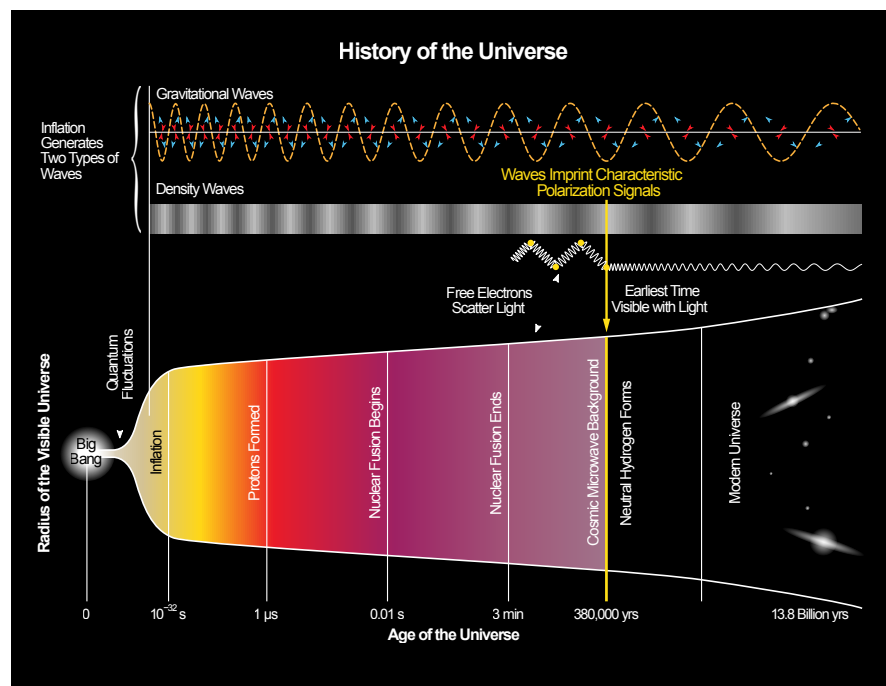


Figure 34.1 Illustration of generation of gravitational waves during the era of cosmic inflation, and how they could be detected through circular polarization imparted on radiation observed in the CMB.

Such explanations for these 3 key problems of the Hot Big Bang model have led to a broad (though not universal) consensus that some form of cosmic inflation did occur in the very early universe, though the details of exactly when and how it was initiated remain uncertain.

However there are experiments underway to detect observational signatures of this inflation era. As illustrated in figure 34.1, the quantum fluctuations in the inflation era, which are thought to cause the fluctuations in the CMB, are also predicted to excite gravitational waves, which are like ripples in the very fabric

of space-time. Such gravitational waves can induce a *circular polarization* in the CMB radiation. The indirect detection of cosmological gravitational waves through circular polarization in the CMB would thus represent an important test of general relativity, as well as provide confirmation for, and observational constraints on, the theory of Cosmic Inflation.

In 2014 there were preliminary claims of such a detection from a project called Bicep2, but so far these have not been confirmed or generally accepted. Indeed, it is now generally believed that the inferred circular polarization signature was likely the result of contamination by foreground dust, and not the sought-after signature of gravitational waves generated by cosmic inflation. But there are hopes that with further improvements in instrumentation and data analysis, it might still be possible in the near future to detect this key signature of cosmic inflation.

Appendix A Atomic Energy Levels and Transitions

As a basis for the examination in part II of how various inferred basic properties of stars from part I can be understood in terms of the physics of stellar structure, let us next consider some key physical underpinnings for interpreting observed stellar spectra. Specifically, this section discusses the simple Bohr model of the Hydrogen atom, while the next section reviews the Boltzmann description for excitation and ionization of atoms.

A.1 The Bohr atom

The discretization of atomic energy that leads to spectral lines can be understood semi-quantitatively through the simple Bohr model of the Hydrogen atom. In analogy with planets orbiting the sun, this assumes that electrons of charge $-e$ and mass m_e are in a stable circular orbit around the atomic nucleus (for hydrogen just a single proton) of charge $+e$ whose mass m_p is effectively infinite ($m_p/m_e = 1836 \gg 1$) compared to the electron. The electrostatic attraction between these charges¹ then balances the centrifugal force from the electron's orbital speed v along a circular orbit of radius r ,

$$\frac{e^2}{r^2} = \frac{m_e v^2}{r}. \quad (\text{A.1})$$

In classical physics, this orbit could, much like a planet going around the sun, have any arbitrary radius. But in the microscopic world of atoms and electrons, such classical physics has to be modified – indeed replaced – by *quantum mechanics*². Just as a light wave has its energy quantized into discrete bundles called photons, it turns out that the orbital energy of an electron is also quantized into discrete levels, much like the steps of a staircase. The basic reason stems

¹ The force on the left-side of (A.1) is written here for CGS units, for which r is in cm and the electron charge magnitude is 4.8×10^{-10} statcoulomb (a.k.a. “esu”), where $\text{statcoulomb}^2 = \text{erg cm} = \text{dyne cm}^2$. For MKS units, for which the charge is 1.6×10^{-19} Coulomb, there is an additional proportionality factor $1/4\pi\epsilon_o$, where $\epsilon_o = 8.85 \times 10^{-12}$ Coulomb²/J/m is the “permittivity of free space”. For simplicity, we use the CGS form here.

² In the classic sci-fi flick *Forbidden Planet*, the chief engineer of a spaceship quips, “I’ll bet any quantum mechanic in the space force would give the rest of his life to fool around with this gadget”.

from the fact that, in the ghostly world of quantum mechanics, electrons are themselves not entirely discrete particles, but rather, much like light, can also have a “wavelike” character. In fact any particle with momentum $p = mv$ has an associated “*de Broglie wavelength*” given by

$$\lambda = \frac{h}{mv}, \quad (\text{A.2})$$

where again, h is Planck’s constant.

This wavy fuzziness means an orbiting electron cannot be placed at any precise location, but is somewhat spread along the orbit. But then to avoid “interfering with itself”, integer multiples n of this wavelength should match the orbital circumference $2\pi r$, implying

$$n\lambda = 2\pi r = \frac{nh}{mv}. \quad (\text{A.3})$$

Note that Planck’s constant itself has units of momentum times distance³, which represents an *angular momentum*. So another way to view this is that the electron’s orbital angular momentum $J = mvr$ must likewise be quantized,

$$J = mvr = n\hbar, \quad (\text{A.4})$$

where $\hbar \equiv h/2\pi$ is a standard notation shortcut. The integer index n is known as the *principal quantum number*.

The quantization condition in eqn. (A.3) or (A.4) implies that the orbital radius can only take certain discrete values r_n , numbered by the level n ,

$$r_n = n^2 \frac{\hbar^2}{m_e e^2} = n^2 r_1, \quad (\text{A.5})$$

which for the ground state, $n = 1$, reduces to the “Bohr radius”, $r_1 \approx 0.529 \text{ \AA} = 0.0529 \text{ nm}$. More generally, this implies that most atoms have sizes of a few Angstrom ($1 \text{ \AA} \equiv 0.1 \text{ nm}$).

It is also useful to cast this quantization in terms of the associated orbital *energy*. The total orbital energy is a combination of the *negative potential* energy $U = -e^2/r$, and the *positive kinetic* energy $T = m_e v^2/2$. Using the orbital force balance eqn. (A.1), we find that the total energy is

$$E_n = -\frac{e^2}{2r_n} = -\frac{m_e e^4}{2\hbar^2} \frac{1}{n^2} = \boxed{-\frac{E_1}{n^2} = E_n}, \quad (\text{A.6})$$

where

$$E_1 \equiv \frac{m_e e^4}{2\hbar^2} = \frac{e^2}{2r_1} = 2.2 \times 10^{-11} \text{ erg} = \boxed{13.6 \text{ eV} = E_1} \quad (\text{A.7})$$

denotes the ionization (a.k.a. binding) energy of Hydrogen from the ground state

³ Or also, energy \times time, which when used with Heisenberg’s Uncertainty Principle $\Delta E \Delta t \gtrsim \hbar$, will lead us to conclude that an atomic state with finite lifetime t_{life} must have a finite width or “fuzziness” in its energy $\Delta E \sim \hbar/t_{life}$. This leads to what is known as “natural broadening” of spectral lines.

(with $n = 1$). Figure A.1 gives a schematic rendition of the energy levels of Hydrogen, measured in *electron Volts* (eV), which is the energy gained when a charge of one electron falls through an electrical potential of one volt.

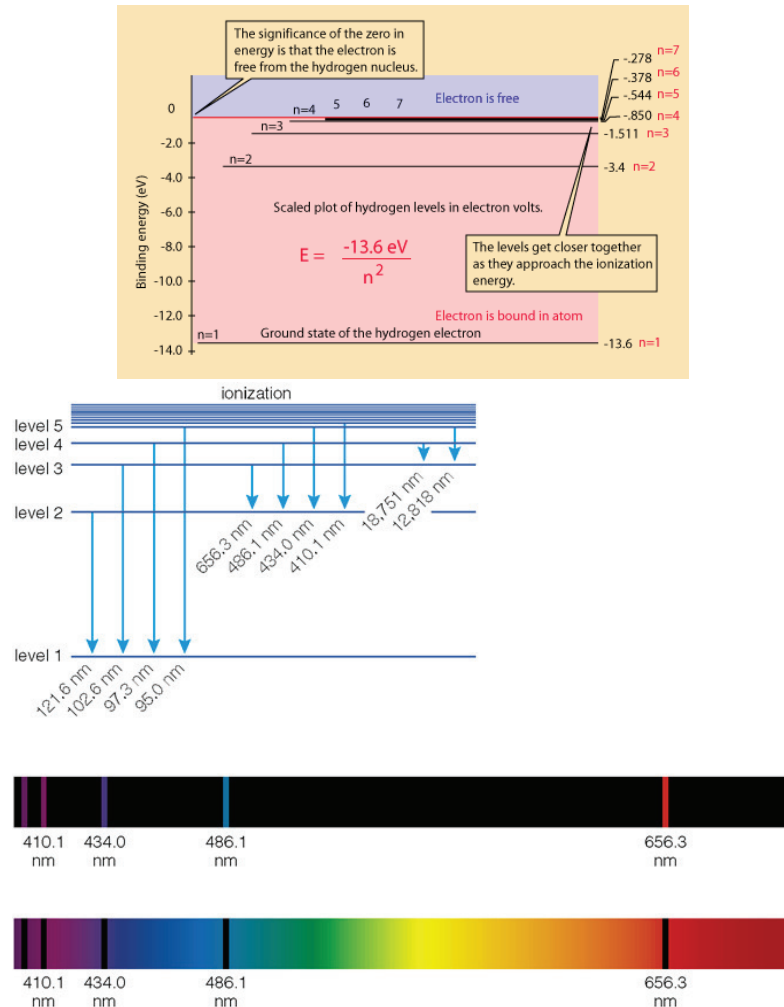


Figure A.1 Top: The energy levels of the Hydrogen atom. The figure is taken from <http://hyperphysics.phy-astr.gsu.edu/hbase/hyde.html#c3>. Middle: Illustration of how the downward transitions between energy levels of a hydrogen atom give rise to emission at discrete wavelengths of a radiative spectrum. Bottom: The corresponding absorption line spectrum at the same characteristic wavelengths, resulting from absorption of a background continuum source of light that then induces *upward* transitions between the same energy levels.

A.2 Emission vs. Absorption line spectra

When an electron changes from one level with quantum number m to another with quantum number n , then the associated change in energy is

$$\Delta E_{mn} = E_1 \left(\frac{1}{n^2} - \frac{1}{m^2} \right) = \boxed{13.6 \text{ eV} \left(\frac{1}{n^2} - \frac{1}{m^2} \right)}. \quad (\text{A.8})$$

If $m > n$ in eqn. (A.8), this represents a positive energy, $\Delta E_{mn} > 0$, which can be *emitted* as a photon of just that energy $h\nu = \Delta E_{mn}$. Conversely, if $m < n$, we have $\Delta E_{mn} < 0$, implying that energy must be supplied externally, for example by *absorption* of a photon of just the right energy, $h\nu = -\Delta E_{mn}$. These processes are called “bound-bound” emission and absorption, because they involve transitions between two bound levels of electrons in an atom.

Bound-bound absorption is the basic process responsible for the absorption line spectrum seen from the surface of most stars. As illustrated in the right panel of figure 6.2, the relatively cool atoms near the surface of the star absorb the light from the underlying layers.

On the other hand, for gas in interstellar space, the atoms are generally viewed against a dark background, instead of the bright back-lighting of a star. If the gas is dense and hot enough that collisions among the atoms occur with enough frequency and enough energy to excite the bound electrons in the atoms to some level above the ground state, then the subsequent *spontaneous decay* to some lower level will emit photons, and so result in an *emission-line spectrum*.

Recall again that figure 6.2 illustrates the basic processes for production of emission and absorption line spectra in both the laboratory and astrophysics.

A.3 Line wavelengths for term series

Instead of photon energy, light is more commonly characterized by its wavelength $\lambda = c/\nu = hc/E$. Using this conversion in eqn. (A.8), we find the wavelength of a photon emitted by transition from a level m to a lower level n is

$$\lambda_{mn} = \frac{\lambda_1}{\frac{1}{n^2} - \frac{1}{m^2}} = \frac{912 \text{ \AA}}{\frac{1}{n^2} - \frac{1}{m^2}}, \quad (\text{A.9})$$

where

$$\lambda_1 \equiv \frac{hc}{E_1} = \frac{h^3 c}{2\pi^2 m_e e^4} = 91.2 \text{ nm} = 912 \text{ \AA} \quad (\text{A.10})$$

is the wavelength at what is known as the *Lyman limit*, corresponding to a transition to the ground state $n = 1$ from an arbitrarily high bound level with $m \rightarrow \infty$. Of course, transitions from a lower level m to a higher level n require *absorption* of a photon, with the wavelength now given by the absolute value of eqn. (A.9).

The lower level of a transition defines a *series* of line wavelengths for transitions from all higher levels. For example, the *Lyman series* represents all transitions to/from the ground state $n = 1$. Within each series, the transitions are denoted in sequence by a lower case greek letter, e.g. $\lambda_{21} = (4/3) 912 = 1216 \text{ \AA}$ is called Lyman- α , while $\lambda_{31} = (9/8) 912 = 1026 \text{ \AA}$ is called Lyman- β , etc. The Lyman series transitions all fall in the ultraviolet (UV) part of the spectrum, which due to UV absorption by the earth's atmosphere is generally not possible to observe from ground-based observatories.

More accessible is the *Balmer series*, for transitions between $n = 2$ and higher levels with $m = 3, 4$, etc., which are conventionally denoted $H\alpha$, $H\beta$, etc. These transitions are pretty well positioned in the middle of the visible, ranging from $\lambda_{32} = 6566 \text{ \AA}$ for $H\alpha$ to $\lambda_{\infty 2} = 3648 \text{ \AA}$ for the *Balmer limit*.

The Paschen series, with lower level $n = 3$, is generally in the InfraRed (IR) part of the spectrum. Still higher series are at even longer wavelengths.

A.4 Questions and Exercises

Quick Questions 1:

- Compute the wavelengths (in nm) for Paschen- α λ_{43} and the Paschen limit $\lambda_{\infty 3}$.
- What are the associated changes in energy (in eV), ΔE_{43} and $\Delta E_{\infty 3}$.

Exercise 1: For an electron and proton that are initially a distance r apart, show that the energy needed to separate them to an arbitrarily large distance is given by $U(r) \equiv -e^2/r$. Use the resulting potential energy $U(r)$ together with the orbital kinetic energy $T = m_e v^2/2$ to derive the expressions in eqn. (A.6) for the total energy $E = U + T$.

Exercise 2: Confirm the validity of eqn. (A.6) by using eqn. (A.1) to show that $E = U/2 = -T$, where U , T , E are the potential, kinetic, and total energy of an orbiting electron. (Note: this result is sometimes referred to as a corollary of the Virial Theorem for bound systems, which is discussed elsewhere in these notes.)

Appendix B Equilibrium Excitation and Ionization Balance

B.1 Boltzmann equation

A key issue for forming a star's absorption spectrum is the balance of processes that excite and de-excite the various energy levels of the atoms. In addition to the photon absorption and emission processes discussed above, atoms can also be excited or de-excited by collisions with other atoms. Since the rate and energy of collisions depends on the gas temperature, the shuffling among the different energy levels also depends sensitively on the temperature.

Under a condition called *thermodynamic equilibrium*, the population of electrons gets all mixed up; then if these levels were all equal in energy, the numbers in each level i would just be proportional to the number of quantum mechanical states, g_i , associated with the orbital and spin state of the electrons in that level¹. But between a lower level i and upper level j with an energy difference ΔE_{ij} , the relative population is also weighted by an exponential term called the *Boltzmann factor*,

$$\frac{n_j}{n_i} = \frac{g_j}{g_i} e^{-\Delta E_{ij}/kT}, \quad (\text{B.1})$$

where $k = 1.38 \times 10^{-16} \text{erg/K}$ is known as Boltzmann's constant. (Also, since energy levels are typically given in electron volt, it is convenient to note that $1 \text{ eV/k} = 1.16 \times 10^4 \text{ K}$.) At low temperature, with the thermal energy much less than the energy difference, $kT \ll \Delta E_{ij}$, there are relatively very few atoms in the more excited level j , $n_j/n_i \rightarrow 0$. Conversely, at very high temperature, with the thermal energy much greater than the energy difference, $kT \gg \Delta E_{ij}$, the ratio just becomes set by the statistical weights, $n_j/n_i \rightarrow g_j/g_i$.

As the population in excited levels increases with increased temperature, there are thus more and more atoms able to emit photons, once these excited states spontaneously decay to some lower level. This leads to an increased *emission* of the associated line transitions.

On the other hand, at lower temperature, the population balance shifts to lower levels. So when these cool atoms are illuminated by continuum light from hot layers, there is a net *absorption* of photons at the relevant line wavelengths, leading to a line-absorption spectrum.

¹ These orbital and spin states are denoted by quantum mechanical numbers ℓ and m , which thus supplement the principal quantum number n .

B.2 Saha equation for ionization equilibrium

At high temperatures, the energy of collisions can become sufficient to overcome the full binding energy of the atom, allowing the electron to become free, and thus making the atom an *ion*, with a net positive charge. For atoms with more than a single proton, this process of *ionization* can continue through multiple stages up to the number of protons, at which point it is completely stripped of electrons. Between an ionization stage i and the next ionization stage $i + 1$, the exchange for any element X can be written as

$$X_{i+1} \leftrightarrow X_i + e^- . \quad (\text{B.2})$$

In thermodynamic equilibrium, there develops a statistical balance between the neighboring ionization stages that is quite analogous to the Boltzmann equilibrium for bound levels given in eqn. (B.1). But now the ionized states consist of both ions, with many discrete energy levels, and free electrons. The number of *bound* states of an ion in ionization stage i is now given by something called the *partition function*, which we will again write as g_i . But to write the equilibrium balance, we now need also to find an expression for the number of states available to the *free* electron.

For this we return again to the concept of the de Broglie wavelength, writing this now for an electron with thermal energy kT . Using the relation $p^2/2m_e = \pi kT$ between momentum and thermal energy, the thermal de Broglie wavelength is

$$\Lambda = \frac{h}{p} = \frac{h}{\sqrt{2\pi m_e kT}} . \quad (\text{B.3})$$

For each of the two electron spins, the total number of free-electron states available per unit volume is $2/\Lambda^3$. For electron number density n_e , this then implies there are $2/n_e \Lambda^3$ states for each free electron.

Using this, we can then describe the ionization balance between neighboring stages i and $i + 1$ through the *Saha-Boltzmann equation*,

$$\frac{n(X_{i+1})}{n(X_i)} = \frac{g_{i+1}}{g_i} \left(\frac{2}{n_e \Lambda^3} \right) e^{-\Delta E_i/kT} , \quad (\text{B.4})$$

where ΔE_i is the ionization energy from stage i , and n_e is the free electron number density. The g_i now represent what's known as the "partition function", which characterizes the total number of bound states available for each ionization stage i ; the large (and formally even divergent!) number of bound states can make it difficult to compute the partition functions g_i , but for Hydrogen under conditions in stellar envelopes, one obtains a typical partition ratio $g_1/g_0 \approx 10^{-3}$.

Throughout a normal star, the electron state factor in parentheses is typically a huge number². For example, for conditions in a stellar atmosphere, it is typically

² As discussed later, it only becomes order unity in very compressed conditions, like in the interior of a white dwarf star, which is thus said to be *electron degenerate*; see sections 16 and 17 of part II.

of order 10^{10} . This large number of states acts like a kind of “attractor” for the ionized state. It means the numbers in the more vs. less ionized states can be comparable even when the exponential Boltzman factor is very small, with a thermal energy that is well below the ionization energy, i.e. $kT \approx \Delta E_i/10$.

For example, hydrogen in a stellar atmosphere typically starts to become ionized at a temperature of about $T \approx 10^4 K$, even though the thermal energy is only $kT \approx 0.86$ eV, and thus much less than the hydrogen ionization energy $E_i = 13.6$ eV, implying a Boltzman factor $e^{-13.6/0.86} = 1.4 \times 10^{-7}$. For a partition ratio $g_1/g_0 \approx 10^{-3}$, we thus obtain roughly equal fractions of Hydrogen in neutral and ionized states at modest temperature of just $T \approx 10^4 K$.

B.3 Questions and Exercises

Quick Question 1: The $n = 2$ level of Hydrogen has $g_2 = 8$ states, while the ground level has just $g_1 = 2$ states. Using the energy difference ΔE_{21} from the Bohr atom, compute the Boltzmann equilibrium number ratio n_2/n_1 of electrons in these levels for a temperature $T = 100,000$ K.

Exercise 1: For a medium of pure hydrogen with total number density $n_H = 10^{10} \text{ cm}^{-3}$, compute the temperature T for unit number ratio of $n_0/n_1 = 1$ for neutral/ionized Hydrogen, assuming a ratio $g_0/g_1 = 2$ for the neutral/ionized states.

Appendix C Atomic origins of opacity

For solid objects in our everyday world, the interaction with light depends on the object's physical projected area, which is the source of the above concept of a “cross section”. But as noted in §12.3, for interstellar dust with sizes become comparable to the wavelength of light, the effective cross section can depend on this wavelength, and so differ from the projected geometric area.

For atoms, ions and electrons that make up a gaseous object like a star, the effective cross sections for interaction with light can be even more sensitive to the details. But generally because light is an Electro-Magnetic (EM) wave, at the atomic level its fundamental interaction with matter occurs through the variable acceleration of charged particles by the varying electric field in the wave. As the lightest common charged particle, electrons are most easily accelerated, and thus are generally key in setting the interaction cross section. The simplest example is that of an isolated free electron, so let's begin by examining its interaction cross section and opacity.

C.1 Thomson cross-section and opacity for free electron scattering

As illustrated in the top left panel of figure C.1, when a passing EM wave causes a free electron to oscillate, it generates a wiggle in the electron's own electric field, which then propagates away – at the speed of light – as a new EM wave in a new direction. Because an isolated electron has no way to store both the energy and momentum of the incoming light, it cannot by itself absorb the photon, and so instead simply scatters, or redirects it. The overall process is called “*Thomson scattering*”.

For such free electrons, the associated Thomson cross section can actually be accurately computed using the classical theory of electromagnetism. Intuitively, the scaling can be roughly understood in terms of the so-called “classical electron radius” $r_e = e^2/m_e c^2$, which is just the radius at which the electron's electrostatic self-energy e^2/r_e equals the electron's rest-mass energy $m_e c^2$. In these terms, the Thomson cross section for free-electron scattering is just a factor¹ 8/3 times greater than the projected area of a sphere with the classical electron

¹ This factor 8/3 comes from detailed classical calculations, and is not easy to understand in simple intuitive terms.

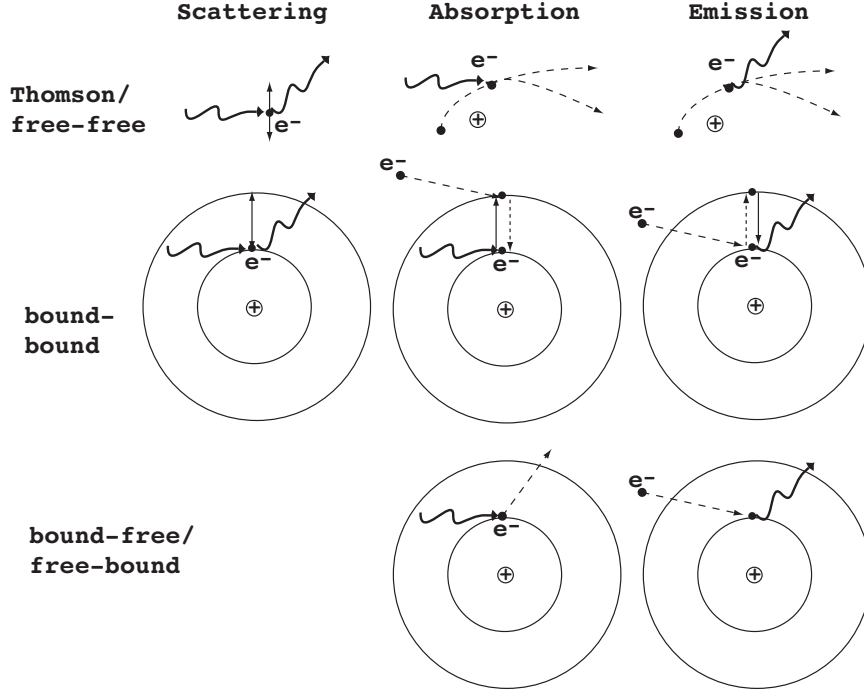


Figure C.1 Illustration of the free-electron and bound-electron processes that lead to scattering, absorption, and emission of photons.

radius,

$$\sigma_{Th} = \frac{8}{3} \pi r_e^2 = \frac{8}{3} \frac{\pi e^4}{m_e^2 c^4} = 0.66 \times 10^{-24} \text{ cm}^2. \quad (\text{C.1})$$

For stellar material to have an overall neutrality in electric charge, even free electrons must still be associated with corresponding positively charged ions, which have much greater mass. Defining then a mean mass per free electron μ_e , we can also define an electron scattering opacity $\kappa_e \equiv \sigma_{Th}/\mu_e$. Ionized hydrogen gives one proton mass m_p per electron, but for fully ionized Helium (and indeed for most all heavier ions), there are two nucleon masses (one proton and one neutron, $m_p + m_n \approx 2m_p$) for each electron. For ionized stellar material with hydrogen mass fraction X , we thus have $\mu_e = 2m_p/(1 + X)$, which then gives for the opacity,

$$\kappa_e \equiv \frac{\sigma_{Th}}{\mu_e} = 0.2 (1 + X) \text{ cm}^2/\text{g} = 0.34 \text{ cm}^2/\text{g}, \quad (\text{C.2})$$

where the last equality assumes a “standard” solar Hydrogen mass fraction $X = 0.72$.

C.2 Atomic absorption and emission: free-free, bound-bound, bound-free

When electrons are bound to atoms or ions, or even just nearby ions, then the combination of the electron and atom/ion can lead to true *absorption* of a photon of light. As shown in the center top row of figure C.1, for free electrons near ions, the shift in the electron trajectory as it passes an ion can now absorb a photon's energy, a process called *free-free absorption*. The right top panel shows that the inverse process can actually produce a photon, and so is called *free-free emission*.

The second row illustrates *bound-bound* processes, involving up/down jumps of electrons between two bound energy levels of atom, with associated absorption/emission of photon (middle and right panel in second row), or indeed, a scattering if the absorption is quickly followed by a reemission of a photon with the same energy, but in a different direction (left panel, second row).

These bound-bound processes only work with photons with just the right energy to match the difference in energy levels, and so lead to the spectral line absorption or emission discussed earlier. But for those “just right” photons, the interaction cross section (leading to the opacity) can be much much higher than for Thomson scattering or free-free absorption, because in effect it is a kind of “resonance” interaction. An everyday analogy is blowing into a whistle vs. just into open air. In open air, you get a weak white noise sound, made up of a range of sound frequencies/wavelengths. With a whistle, the sound is loud and has a distinct pitch, representing a resonance oscillation at some well-defined frequency/wavelength.

The third row illustrates the *bound-free* processes associated with a photon absorption that causes an atom or ion to become (further) ionized by kicking off its electron. As with electron scattering or even free-free absorption, it is a continuum (vs. line) process, though it does now require that the photons have a energy equal to or greater than the ionization energy for that atom or ion. Its interaction cross-section can be significantly higher than electron scattering or free-free absorption, but is generally not as strong as for bound-bound processes that lead to lines.

The cross sections, and corresponding opacities, associated with these electron+ion/atom processes are much more complicated than for free electrons, and so are difficult to cast in the kind of simple formula given in eqn. C.2 for Thomson electron scattering opacity. But often bound-free and free-free opacities are taken to follow a so-called “Kramer’s opacity”, for which

$$\kappa_{kr} \sim \rho T^{-7/2} \sim (P_{gas}/P_{rad}) T^{-1/2}. \quad (C.3)$$

As discussed further below, often in stellar interiors the ratio of gas to radiation pressure is nearly constant, so that opacity decreases only weakly (as $1/\sqrt{T}$) with the increasing temperature of the interior.

A simple rough rule of thumb is that, outside of ionization zones where bound-free absorption can substantially enhance the overall opacity, stellar interiors

typically have opacities that are some modest factor few times the simple electron scattering opacity in (C.2), i.e., with a characteristic CGS value of order unity, $\kappa \approx 1 \text{ cm}^2/\text{g}$.

C.3 Questions and Exercises

Quick Question 1:

- a. Seen standing up, what is the cross section (in cm^2) of a person with height 1.8 m and width 0.5 m?
- b. If this person has a mass of 60 kg, what is his/her “opacity” $\kappa = \sigma/m$, in cm^2/g ?
- c. How does this compare with the typical opacity of stellar material?

Appendix D Radiative Transfer

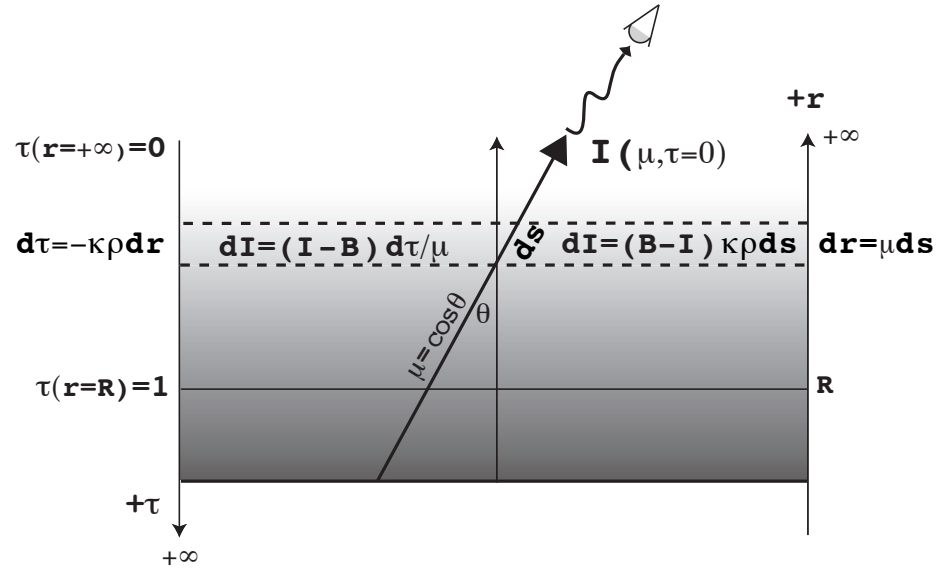


Figure D.1 Emergent intensity from a semi-infinite, planar atmosphere. Along a direction \hat{s} that has projection $\mu = \hat{s} \cdot \hat{r} = \cos \theta$ to the local vertical (radial) direction \hat{r} , the change in intensity in each differential layer dr depends on thermal emission of radiation by the local Planck minus the absorption of local intensity I , multiplied by the projected change in optical depth $-d\tau/\mu = \kappa\rho ds$ along the path segment ds .

D.1 Absorption and thermal emission in a stellar atmosphere

As noted above (§15.2), the atmospheric transition between interior and empty space occurs over a quite narrow layer, a few scale heights H in extent, which typically amounts to about a thousandth of the stellar radius (cf. eqn. 15.5). At any given location on the spherical stellar surface, the transport of radiation through this atmosphere can be modeled by treating it as a nearly *planar* layer, as illustrated in figure D.1.

To quantify this atmospheric transition between random-walk diffusion of the

deep interior to free-streaming away from the stellar surface, we must now solve a differential equation that accounts for the competition between the reduction in intensity due to absorption vs. the production of intensity due to the local thermal *emission* $B(\tau)$. As illustrated in figure D.1, consider a planar atmosphere with an arbitrarily large optical depth (at bottom) seen from an observer at optical depth zero (at top) who looks along a direction \hat{s} that has a projection¹ $\mu = \cos \theta$ to the local vertical (radial) direction \hat{r} . The change in intensity in each differential layer dr depends on thermal emission of radiation by the local Planck function B minus the absorption of local intensity I , multiplied by the projected change in optical depth $-d\tau/\mu = \kappa \rho ds$ along the path segment ds . This leads to an “equation of radiative transfer”,

$$\mu \frac{dI(\mu, \tau)}{d\tau} = I(\mu, \tau) - B(\tau), \quad (\text{D.1})$$

where the radial optical depth integral is now defined *from* a distant observer at $r \rightarrow \infty$,

$$\tau(r) \equiv \int_r^\infty \kappa \rho dr', \quad (\text{D.2})$$

which thus places the observer at $\tau(r \rightarrow \infty) = 0$.

D.2 The Eddington-Barbier relation for emergent intensity

Eqn. (D.1) is a linear, first-order differential equation. As discussed in the exercise below, by using integrating factors, it can be converted to a formal integral solution for the emergent intensity seen by an external observer viewing the atmosphere along a projection μ with the local radius,

$$I(\mu, \tau = 0) = \int_0^\infty B(\tau) e^{-\tau/\mu} d\tau / \mu \approx B(\tau = \mu). \quad (\text{D.3})$$

The latter approximation here assumes the Planck function is roughly a linear function of optical depth near the star’s surface, $B(\tau) \approx a + b\tau$. This so-called “Eddington-Barbier relation” states that when you peer into an opaque radiating gas, the emergent intensity you perceive is set by the value of the blackbody function at the location of unit optical depth along that ray. This in turn is set by the temperature at that location, providing a more rigorous definition for what we’ve referred to up to now as surface brightness and surface temperature.

An example of this E-B relation comes from the observed “limb darkening” of the solar disk, as illustrated by the visible light picture of the sun in fig. D.2. Because the line of sight looking at the center is more directly radial to the sun’s local surface, one can see into a deeper, hotter layer than from the more

¹ This standard notation using μ for direction cosine here should not be confused with the notation in the previous sections that use μ for molecular weight.

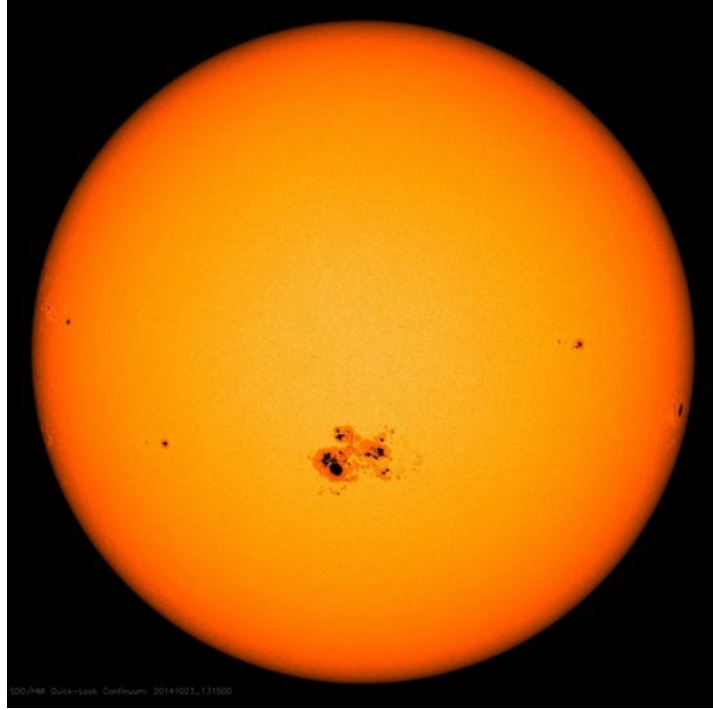


Figure D.2 Visible light picture of the solar disk, showing the center to limb darkening of the surface brightness. The central lower group of dark sunspots are regions where solar magnetic storms have inhibited the upward convective transport of heat, leading to a locally cooler and thus darker surface.

oblique angle when viewing toward the edge or “limb” of the solar disk. This makes the disk appear brightest at the center, and darker as the view moves toward the solar limb². The observed variation from center to limb thus provides a diagnostic of the *temperature gradient* in the sun’s surface layers.

Since stars are too far away to resolve their angular size, we can’t observe their emergent intensity $I(\mu, 0)$, but we can observe the flux $F(r) = L/4\pi r^2$ associated with the total luminosity $L = 4\pi R^2 \sigma_{sb} T_{\text{eff}}^4$. The emergent *surface* flux $F_* = L/4\pi R^2$ is obtained by integrating $\mu I(\mu, 0)$ over the 2π solid angle for the

² Of course, the brightness of the sun means we need special filters to see this effect. One should *never* look at the sun with the naked eye.

hemisphere open to empty space, giving

$$\begin{aligned}
 F_* &\equiv 2\pi \int_0^1 \mu I(0, \mu) d\mu \\
 &\approx 2\pi \int_0^1 \mu B(\tau = \mu) d\mu \\
 &\approx 2\pi \int_0^1 \mu (a + b\mu) d\mu \\
 &= \pi B(\tau = 2/3) \\
 &= \boxed{\sigma_{sb} T^4(\tau = 2/3)}, \tag{D.4}
 \end{aligned}$$

where the third equality assumes the Planck function near the surface can be approximated as a linear function of optical depth, $B(\tau) \approx a + b\tau$.

Comparison of the final form of (D.4) with the simple discussion of surface flux in part I shows that we can identify what we've been calling the stellar “surface” as the layer where the optical depth $\tau(R) \equiv 2/3$, with the “surface temperature” likewise just the temperature at this layer.

Stars are not really black-bodies, but it is convenient to *define* a star's “effective temperature” T_{eff} as the blackbody temperature that would give the star's inferred surface flux $F_* = L/4\pi R^2$. From (D.4), we see that we can associate this effective temperature with the surface temperature at optical depth $2/3$, $T_{\text{eff}} = T(\tau = 2/3)$.

D.3 Questions and Exercises

Exercise 1: Derive the integral solution (D.3) from the differential equation (D.1), assuming a semi-infinite atmosphere that extends to large depths $\tau \rightarrow \infty$. *Hint:* First multiply eqn. (D.1) by an integrating factor $e^{-\tau/\mu}$, and use this to write the change in intensity in terms of a full differential. Then carry out the integral from the observer at $\tau = 0$ to some finite depth τ where the intensity is taken to have a given value $I(\tau, \mu)$. Finally take the limit $\tau \rightarrow \infty$ to obtain (D.3).

Exercise 2: If thermal emission from the Planck function is a linear function of radial optical depth $B(\tau) = a + b\tau$, explicitly do the integration in (D.3) to derive the Eddington-Barbier relation for emergent intensity $I(\mu, 0) = B(\tau = \mu)$.